UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

MASTER'S THESIS

Limitations and challenges of unsupervised cross-lingual pre-training

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging
Academic Course 2020/2021

Martín Quesada Zaragoza

Adviser:
Francisco Casacuberta Nolla

Abstract / Resum / Resumen

Abstract

Cross-lingual alignment methods for monolingual language representations have received notable research attention in the past few years due to their capacity to induce bilingual alignments with little or no supervision signals. However, their use in machine translation pre-training, a function that monolingual models excel at, and which should benefit from cross-lingual information, remains limited. This work tries to shed light on the effects of some of the factors that play a role in cross-lingual representations and pre-training strategies, with the hope that it can help guide future endeavors in the field. To this end, the survey studies the two main components that constitute cross-lingual pre-training: cross-lingual mappings and their pre-training integration. The former are explored through some widely known fully unsupervised cross-lingual methods, which rely on distributional similarities between languages. Consequently, they are a great basis upon which to consider the effects of language similarity on both cross-mapping techniques and the representation spaces over which they operate. In pre-training integration, cross-lingual representation spaces are used to pre-train a neural machine translation models, which are compared against techniques that employ independent monolingual spaces. The results show that weakly-supervised cross-lingual methods are remarkably effective at inducing alignment even for distant languages and they benefit noticeably from subword information. However, the effect of cross-linguality in pre-training is diminished due to difficulties in maintaining the structure of the projection during training, and the limited influence that pre-training itself has in the supervised model.

Keywords

cross-lingual embeddings, bilingual lexicon induction, machine translation pre-training, cross-lingual pre-training, neural machine translation

Resum

Els mètodes d'alineament croslingüe per a representacions monolingües del llenguatge han sigut objecte d'un interés notable en el camp de processament del llenguatge natural durant els últims anys, en gran manera a causa de la capacitat que aquests tenen per a general alineaments entre llengües utilitzant poca o nul·la informació paral·lela. No obstant això, el seu ús en tècniques de preentrenament de models de traducció automàtica, un paper en el qual els models monolingües són particularment reeixits, i que hauria de beneficiar-se de la informació croslingüe obtinguda, continua sent limitat. Aquesta proposta intenta aportar una mica de llum sobre els efectes d'alguns dels factors que afecten les representacions croslingües i les estratègies de preentrenament, amb l'esperança que puga ajudar a futures investigacions en aquest camp. Per a això, aquest treball estudia els dos components principals que constitueixen el preentrenament croslingüe: els alineaments croslingües i la integració dels mateixos com a models de preentrenament. Els primers són explorats a través de diversos mètodes croslingües no supervisats àmpliament coneguts, que empren principalment similituds distribucionals per a trobar un alineament satisfactori entre llenguatges. A causa d'això, resulten un interessant terreny de proves en el qual analitzar els efectes de la similitud entre llenguatges sobre tant les tècniques d'alineament croslingüe com els espais de representació sobre els quals operen. En en apartat d'integració en preentrenament, els espais de representació croslingües són utilitzats per a preentrenar models de traducció automàtica, els quals són comparats contra esquemes que empren espais de representació independents. Els resultats mostren que els mètodes croslingües amb supervisió feble són remarcablement efectius a l'hora de generar alineaments fins i tot per a parelles de llenguatges molt diferents, i es beneficien notablement de la informació a nivell de subparaula. No obstant això, l'efecte de l'alineament croslingüe en el preentrenament és reduït a causa de les dificulteu de mantindre l'estructura de la projecció durant l'entrenament, així com per la limitada influència que el propi preentrenament té sobre el model supervisat.

Paraules clau

embeddings croslingües, inducció de lèxic bilingüe, preentrenament en traducció automàtica, preentrenament croslingüe, traducció automàtica neuronal

Resumen

Los métodos de alineamiento croslingüe para representaciones monolingües del lenguaje han sido objeto de un interés notable en el campo de procesamiento del lenguaje natural durante los últimos años, en gran medida debido a la capacidad que estos tienen para general alineamientos entre lenguas utilizando poca o nula información paralela. Sin embargo, su uso en técnicas de preentrenamiento de modelos de traducción automática, un papel en el que los modelos monolingües son particularmente exitosos, y que debería beneficiarse de la información croslingüe obtenida, sigue siendo limitado. Esta propuesta intenta aportar algo de luz sobre los efectos de algunos de los factores que afectan a las representaciones croslingües y las estrategias de preentrenamiento, con la esperanza de que pueda ayudar a futuras investigaciones en este campo. Para ello, este trabajo estudia los dos componentes principales que constituyen el preentrenamiento croslingüe: los alineamientos croslingües y la integración de los mismos como modelos de preentrenamiento. Los primeros son explorados a través de varios métodos croslingües no supervisados ampliamente conocidos, que emplean principalmente similaridades distribucionales para encontrar un alineamiento satisfactorio entre lenguajes. Debido a esto, resultan un interesante terreno de pruebas en el que analizar los efectos de la similaridad entre lenguajes sobre tanto las técnicas de alineamiento croslingüe como los espacios de representación sobre los que operan. En en apartado de integración en preentrenamiento, los espacios de representación croslingües son utilizados para preentrenar modelos de traducción automática, los cuales son comparados contra esquemas que emplean espacios de representación independientes. Los resultados muestran que los métodos croslingües con supervisión débil son remarcablemente efectivos a la hora de generar alineamientos incluso para parejas de lenguajes muy diferentes, y se benefician notablemente de la información a nivel de subpalabra. Sin embargo, el efecto del alineamiento croslingüe en el preentrenamiento es reducido debido a las dificultad de mantener la estructura de la proyección durante el entrenamiento, así como por la limitada influencia que el propio preentrenamiento tiene sobre el modelo supervisado.

Palabras clave

embeddings croslingües, inducción de léxico bilingüe, preentrenamiento en traducción automática, preentrenamiento croslingüe, traducción automática neuronal

Contents

Chapter 1
Introduction

Machine translation has experienced unprecedented growth in the last decade, a time during which the improvements in translation quality have propelled it to become incredibly wide-spread in short-form internet content, which constitutes most of multimedia and written content worldwide if we account for social media. Still, professional translation services remain very much essential when it comes to commercial and otherwise long-form content, a frontier that machine translation is still far from crossing, and that it might not be able to traverse with our current approach. This period has also signified a transition away from the so called statistical machine translation models (SMT), which presents separated translation, reordering and language model components, and into neural machine translation (NMT) architectures, particularly the attention-based encoder-decoder architectures (Bahdanau, Cho, and Bengio, 2014), which constitute end-to-end solutions. Although modern deep neural translation models heavily rely on the bases established by SMT and prior machine translation technology, they present a much larger number of parameters to tune and are overall more complex systems that have the capacity to surpass previous machine translation benchmarks when provided with large training corpora (Koehn and Knowles, 2017).

However, the cost of training these multi-layered neural networks increases multiple orders of magnitude when compared to SMT models. As a result, great computing power and vast amounts of bilingual training corpora are needed in order for deep neural networks to be effective (Koehn and Knowles, 2017). Such collections of data are usually presented as parallel texts aligned at the sentence or word level, which are costly to produce, since they require manual labour at scale from trained professionals. Therefore, the number of suitable corpora available is reduced, even more so for rare languages and specific fields of interest. This means that, for any use case other than creating a general purpose translator for a widely used language, finding adequate datasets tends to be an issue, while building one's own continues to be extremely expensive and time-consuming due to the volume of data required. Additionally, training times are considerable and often prohibitive for modern deep models (Sharir, Peleg, and Shoham, 2020).

Monolingual pre-training methods aim to tackle both the corpus and training time limitations by providing an initial continuous representation or encoding for the incoming and generated words that is trained over monolingual corpora. This facilitates building models that target obscure languages or specific fields – i.e. medicine, legal, finance – , since the pre-training step can be done over monolingual data, of which there is plenty of no matter the language or subject matter. Although in recent times pre-training models have come to represent increasingly bigger sections of the neural network (Devlin et al., 2019), they were initially conceived as a way to independently train a few of the initial and final layers of the model – relative to the source and target language respectively – , known as the embedding layers (Bengio et al., 2003).

This concept of "embedding" words and sentences into a continuous vector space has been inherited from the greater natural language processing (NLP) field, where numeric representations of words that encode meaning based on context similarity have long been used (Naseem et al., 2020). Originally the most common models were sparse matrices based on term co-occurrence or term frequency—inverse document frequency, among others. These high-dimension representations have their counterparts in dense embeddings, which have been become especially popular with the advent of neural embeddings. Neural embeddings use a shallow neural network to generate a dense vector space based on the context in which a word appears. This is very much part of what a neural network model that receives text as input is doing, which is the reason why neural embeddings also started to be used as monolingual pre-training for larger translation or classification models.

Another interesting property of dense embeddings is the distributional similarities that exist between embeddings, even when they are trained using different languages. The authors of one of the earliest neural word embeddings to enjoy widespread use, Word2Vec (Mikolov, Chen, et al., 2013), already noted this characteristic and published on it in Mikolov, Le, and Sutskever (2013), just a few months after unveiling the new embedding model. This initial cross-lingual method was used to automate the process of generating and extending dictionaries and phrase tables, which are vital in SMT models. Since then, many other approaches to cross-lingual alignment for monolingual word embeddings have appeared, with a similar focus on automatizing the creation of bilingual dictionaries.

Most of these methods are semi-supervised, meaning that, although they work over monolingual embeddings, they use small collections of parallel texts or dictionaries as seeds for the cross-lingual alignment. However, later research has also given birth to fully unsupervised cross-lingual embeddings (CLE), which do not need any bilingual signal. Unsupervised methods can generate their own bilingual dictionary or parallel corpus by relying on generative adversarial networks (Goodfellow et al., 2014), as in MUSE (Lample, Denoyer, and Ranzato, 2017), or calculating monolingual similarity distribution vectors and directly extracting nearest neightbours for the seed dictionary, like in VecMap (Artetxe, Labaka, and Agirre, 2018), among other possible strategies. Fully unsupervised cross-lingual mappings are very interesting due to them being solely reliant on the distributional similarities of language continuous vector representations. This apparent promise of a future where no hand-crafted bilingual supervision is needed has granted these approaches a decent amount of research attention.

However, semi-supervised cross-lingual pre-training methods, which require very small amounts of parallel corpora – and in some cases only a simple bilingual dictionary for initialization – tend to outperform their unsupervised counterparts (Vulic, Glavas, et al., 2019; Doval et al., 2019; Patra et al., 2019). It is only in situations where no bilingual data exists that unsupervised techniques are preferable. However, since the monolingual embeddings and models required to produce a cross-lingual mapping need to be trained over a considerable volume of monolingual text, a scenario where unsupervised cross-lingual alignments are truly necessary seems unlikely. In the case of languages with extensive written records where resources are plentiful, there is little reason to use a fully unsupervised method rather than one that takes advantage of a small bilingual dataset (Artetxe, Ruder, et al., 2020). Even for low-resource languages for which a reduced number of corpora exist, it is exceedingly probable that some sort of translation dictionary that associates them with a more widely studied language is available. For extinct languages, the remaining texts are usually too brief, meaning that the task is best suited for human experts, who can take advantage of contextual information like archaeological evidence or metalinguistic features. This leads us to the question: are unsupervised cross-lingual models useful at all?

While they might not be the best performing tools in a realistic use case, drawing this sort of hard line between unsupervised and semi-supervised cross-lingual mappings does not make much sense, because they are extremely similar processes. Semi-supervised cross-lingual methods are just forfeiting the generation of a seed bilingual dictionary in favour of one provided by the user. But their remaining steps are analogous to that of unsupervised cross-lingual projection methods: they both use the seed translation dictionary to align the monolingual subspaces in a hypothetical cross-lingual space, and then project said embeddings into the aforementioned cross-lingual shared space. In the case of many unsupervised models, seed generation is actually an adaptation of these previous steps, where some assumption on the structures of the matrices (i.e. isometry) is made in order to obtain an initial set of translation pairs. Given that unsupervised and semi-supervised cross-lingual strategies share so many traits, most improvements over unsupervised methods can be transferred to semi-supervised ones. And unsupervised strategies lend themselves very well to experimentation, since there is no variance derived from the specific characteristics of the provided hand-crafted parallel corpus – because there is no parallel corpus to speak of – . As a result, it is more convenient to design and test a formal solution using unsupervised cross-mapping for experimentation, and move into semi-supervision when the interactions between languages and parameters have been examined.

This independence from corpora also makes unsupervised cross-lingual mappings a great vehicle to explore how different continuous representations may capture distinct features depending on the language that they are trained on. For instance, at first glance different languages seem to produce similar embeddings with comparable distributions of data points. However, unsupervised cross-lingual embeddings often are unable to generate an initial alignment when operating with distant language pairs (Vulic, Glavas, et al., 2019; Doval et al., 2019), which leads to near-zero bilingual lexicon induction (BLI) performance. This can be due to the fact that the isomorphism assumption that most approaches take tends to weaken when using increasingly etymologically distant languages, (Patra et al., 2019). In turn, this property can be explained by the findings in Nakashole and Flauger (2018), who propose that embedding spaces in different languages are linearly equivalent only at local regions, but their global structure is different. Both of the previous publications that help explain the nature of embeddings across languages use cross-lingual embeddings as the basis for their experiments, often focusing on the differences between the unsupervised and semi-supervised methods. Unsupervised cross-lingual mappings are therefore very helpful in bringing to light the degree of similarity between continuous representations of different languages, which can point the way to on how to improve said representation spaces.

1.1 Objectives and work structure

In this work, I intend to provide an overview of some representative unsupervised cross-lingual methods and analyze their interactions with some of the factors that influence these kind of models, such as vector space dimension, language effect or the alignment policy itself. The objective is to showcase the limitations of some of the current cross-lingual pre-training approaches, and what this means for the future of the field in regards to its practicality and value as an area of research.

This work is structured according to the two main components that make cross-lingual pre-training possible: cross-lingual mapping techniques and pre-training integration of said methods. Initially, the motivation and historical context behind these ideas is introduced in chapters 1 and 2. Chapter 1 provides an initial breakdown of the motivation behind this work and the areas that it intends to explore. In chapter 2, a brief description of the history, motivation and current state of cross-lingual research is given so that the reader is familiar with the context in which this work is situated, as well as some prior breakthroughs and relevant techniques that can aid in understanding the experiments performed in this work. After this, a technical description of the cross-mapping and pre-training strategies chosen for experimentation is provided in chapters 3 and 4 respectively, as well as some clues on known limitations and the expected effect of certain factors. Chapter 5 takes on implementation specifics of the proposed experiments, such as the adopted evaluation criteria and the rationale behind parameter tuning for the models used, plus some development results. Next in chapter 6, the results of said experiments are shown and discussed. Finally, some conclusions to the work are offered in chapter 7.

Chapter 2
Related work

Natural language processing has been one of the main areas of interest of computing since the inception of the field. Although initially it was treated as a mere component of early artificial intelligence problems, such as the very well known Turing test (Turing, 1950), NLP began to differentiate itself as more language-specific tasks began to be tackled through computational meanings. From these, it is perhaps machine translation – arguably still the flagship task of natural language processing – which raised the most interest in its practical possibilities. The Georgetown experiment (Hutchins, Dostert, and Garvin, 1955) showcased machine translation of more than sixty Russian sentences into English, and its authors projected exponential progress in the field – a prediction which never came true in the short term, but that sparked great interest in this area of research –. This attempt at machine translation, as most of the natural language processing systems designed until at the time, followed suit with the general artificial intelligence field by relying on symbolic methods, which featured extensive, manually designed rule-based models. Symbolic methods remained the dominant paradigm until the 1980s, when the popularization of machine learning algorithms, in combination with increased available computational power and subsequent decrease in dominance of Chomskyan rule-based systems, laid the ground for statistical NLP methods to become widespread. Statistical NLP relies on models that use statistical inference to perform their tasks, learning their parameters over a given training corpus. Therefore, the approach is based on corpus linguistics rather than on known language rules and feature. The relevance of sentence-based statistical NLP systems has lessened in the last decade in favour of neural networks models. These are in many ways an extension of the classic statistical paradigm, as they are similarly reliant on annotated corpus to train an inference model. However, instead of needing of composite systems interconnected to perform different aspects of an overarching high-level task, they integrate complex end-to-end models with high computational costs but increased performance (Koehn and Knowles, 2017). Statistical methods are still used in some areas due to them needing lower volumes of annotated corpora and presenting some statistical interpretability – which is one of the main limitations of neural networks –, particularly in cascaded approaches for voice recognition (Stentiford and Steer, 1988; Waibel et al., 1991; Povey et al., 2011), although this field is also currently moving towards end-to-end integration using neural networks (Di Gangi, Negri, and Turchi, 2019; Inaguma et al., 2019; Sperber et al., 2019).

Statistical and neural NLP require numeric representations of the text that is given as input to the model, since the models are, in essence, statistical inference machines that are working over an unknown distribution. It is therefore unsurprising that word embeddings (numeric representations of words) appeared and continued to evolve alongside statistical NLP methods. Word representation models can be classified into two main archetypes: non-distributional models and distributional methods. In the former, the representations are usually atomic and there is no notion of similarity between the words, which are just indices in a vocabulary. The latter projects words to a continuous vector space words based on their surrounding information, where words from similar contexts are taken to be semantically related and therefore are situated closer to each other in the representation space. Continuous word vectors appeared as a mean to combine existing knowledge on distributed representations (Hinton, McClelland, and Rumelhart, 1986) with the principles behind distributional semantics, providing much needed semantic context to the required numeric representations of text in NLP tasks. A common approach for distributional methods in relatively simple tasks such as indexing are sparse vectors, like term frequency—inverse document frequency (TF-IDF), term-document and term-term matrices – also known as co-occurrence matrices – . Sparse representations have the drawback of producing spaces of extremely high dimension when working over large vocabularies, so they are not well fit for machine learning tasks, where a large amounts of text are used for training. A possible solution that reduces the dimension of the vector spaces is to apply singular value decomposition (SVD) to any of the sparse matrices in order to reduce the number of rows – the size of the vectors that represent each word in the vocabulary – while preserving the mutual information between columns, that is, the word vectors themselves –. Other approaches to distributional embeddings with reduced dimension also include generative models such as Latent Dirichlet Allocation (LDA). These sort of models are referred to as dense embeddings, as they compress mutual information to fit vector spaces of low dimension. Dense distributional models perform much better over large corpora than their sparse counterparts, although they generally require long training times. This led to the popularization of some non-distributional models, namely simple n-gram models, which are cheap to train and benefit from large volumes of monolingual corpora. The next development in distributional methods would come from their combination with neural networks, which produced the first neural word embeddings.

Although nowadays it is common to refer to all distributional word vector methods as word embeddings, this concept was conceived in association with the models that now are mostly regarded as neural word embeddings. The term word embedding was originally coined in Bengio et al. (2003), where the authors trained the embedding within a neural language model using shared parameters. A major breakthrough in performance would come again five years later in Collobert and Weston (2008), a publication that presented semi-supervised learning applied effectively to a multitask network. But it is arguably the work of Mikolov, Chen, et al. (2013) that launched neural word embeddings to the forefront of the field with Word2Vec, a toolkit which employs a shallow neural network architecture that is able to implicitly capture mutual contextual information within a word vector space. Another architecture that is now widespread in the NLP field is GloVe (Pennington, Socher, and C. Manning, 2014), a model that seeks to capture the ratio of co-occurrence probability of words in a condensed vector from a given term co-occurrence matrix. That is, it tries to capture mutual information explicitly rather than implicitly. Modern neural word embeddings, especially the previously mentioned models, have become popular because they are generally robust, easy to apply and their performance is akin to that of the slower classic distributional methods , which has helped them take the place n-gram models in providing a fast solution for models trained over extensive datasets.

Word2Vec in particular became the catalyst for the group of techniques that concerns this work: cross-lingual embeddings, which aim to learn joint cross-lingual word embedding spaces between different languages. On of the most important findings that launched further research on the topic was Mikolov, Le, and Sutskever (2013), which proposed that it is possible to find structural similarities between any two continuous word embeddings, and demonstrated this phenomenon using the then recent Word2Vec model. The approach described in this publication was a simple linear mapping between the matrices represented by monolingual embeddings, with the objective of creating a shared bilingual vector space. This would be the first of a class of cross-lingual methods known as mapping-based cross-lingual embeddings, which try to learn a linear mapping between monolingual embeddings. Other early methods included this paradigm are that of Faruqui and Dyer (2014), Xing et al. (2015), Vulić and Korhonen (2016), Artetxe, Labaka, and Agirre (2016), and Barone (2016). There are also other categories of cross-lingual embeddings that are not projection-related, such as pseudo-cross-embeddings, which train a word embedding over pseudo-cross-lingual corpora – that is, corpora that include texts of similar context in two different languages, but are not directly parallel to each other – (Xiao and Guo, 2014; Gouws and Søgaard, 2015; Duong et al., 2016). Or cross-lingual training, a family of methods that trains embeddings on a parallel corpus and optimize a cross-lingual constraint, seeking to learn to project monolingual word embeddings into a shared vector space while taking into consideration their shared structural similarities, so that words from different that are semantically equivalent are situated closely together in the projected representation space (Hermann and Blunsom, 2014; Lauly, Boulanger, and Larochelle, 2014; Kociský, Hermann, and Blunsom, 2014). Another group of approaches is joint optimization, which proposes to train models on parallel corpora and jointly optimises a combination of monolingual and cross-lingual losses (Klementiev, Titov, and Bhattarai, 2012; Zou et al., 2013; Luong, Pham, and C. D. Manning, 2015; Gouws, Bengio, and Corrado, 2015). Although some of these methods predate or are contemporaries of the linear mapping proposed in Mikolov, Le, and Sutskever (2013), the general interest sparked by mapping cross-lingual embeddings has brought increased attention into this field of research and accelerated the growth of many of them. One of the reasons for this interest is that more modern mapping-based cross-lingual embeddings, which are sometimes also called projection-based embeddings, have become very popular by outperforming earlier mapping methods and many supervised cross-lingual methods that are dependent on segment-level alignment. As many early mapping-based approaches (Mikolov, Le, and Sutskever, 2013; Faruqui and Dyer, 2014), they require monolingual embeddings, but often also a small parallel corpus or a bilingual seed translation dictionary to perform the initial alignment, sometimes optionally (Lample, Denoyer, and Ranzato, 2017). However, as it is also the case for certain mapping-based embeddings, some do not need any parallel signal at all, as they can also perform the original alignment relying only in estimations based on structural similarities between the monolingual vector spaces (Artetxe, Labaka, and Agirre, 2018; Lample, Denoyer, and Ranzato, 2017). This makes them effective for prototyping in extremely low-resource languages models when no parallel data is available. Moreover, it also makes them uniquely suited to research the effect of the different factors that affect cross-lingual methods of any kind, since they are not affected by the particularities of a parallel corpus and will model consistently similar matrices for the same language given sufficient training data.

Cross-lingual methods have also been successfully applied to neural language models used in deep neural network pre-training. Although cross-lingual embeddings have not been particularly effective and often do not seem to represent a significant improvement over using off-the-shelf monolingual embeddings (Qi et al., 2018), cross-lingual language models have fared better. Neural language models have gained notoriety recently due to how effective they are at pre-training deep neural network models using exclusively monolingual training corpora. Unlike neural network-based word embeddings such as Word2Vec, which is based on a shallow network, neural language models are much deeper and costly to train – they often are deep neural networks themselves –, but have the advantage of pre-training the final model to a much higher degree. Therefore, the amount of parallel data needed to train the model to completion is lower, since only a reduced number of layers need to be added to the output of the neural language model and refined. In the field of machine translation, perhaps the most important neural language model to become widespread is BERT (Devlin et al., 2019), based on the Transformer (Vaswani et al., 2017) neural network architecture and originally designed for multiple natural language processing tasks in mind. The effectiveness of BERT in machine translation (Yang et al., 2020) has motivated multiple derivations of the model with this area in mind, some focused on machine translation, such as BART (M. Lewis et al., 2020) or mBART (Liu et al., 2020). Different cross-lingual methods for these methods have been proposed, some try to create a universal neural language model for all the languages included in a final multilingual system, inducing cross-lingual information through zero-shot transfer (Ji et al., 2020) or cross-mapping all languages together using random aligned substitution (Lin et al., 2021). In contrast, Lample and Conneau (2019) create a cross-lingual neural language model by training a masked language model (Devlin et al., 2019) using a shared vocabulary between languages and subsampling frequent outputs as per Mikolov, Sutskever, et al. (2013). Ren et al. (2019) refine this masked language model (MLM) by using an initial n-gram translation table inferred unadvisedly, and introducing an explicit cross-language training objective, creating a cross-lingual masked language model (CMLM). Recent research in Wang and Zhao (2021) has obtained top-of-the-line results by using a large-scale CMLM and training the final supervised model using a joint optimization objective (Sun et al., 2019) that is intended to maintain the original distribution of the CMLM as much as possible while maximizing the translation performance of the global model.

Chapter 3

Unsupervised cross-lingual models

3.1 Projection-based cross-lingual embeddings

Both of the cross-lingual embedding methods considered in this work are projection-based cross-lingual embeddings. Projection-based CLE try to produce an alignment between monolingual word embeddings, subsequently projecting them into a common representation space that facilitates a direct mapping between the distribution of both embeddings, using as a guide for the alignment a number of translation pairs that serve as a seed dictionary. However, in some cases these translation pairs are estimated by the cross-lingual method itself. This capacity to generate a common vector space of aligned embeddings with no bilingual signal to speak of defines fully unsupervised cross-lingual embeddings, which is the subset of methods in which all the methods considered in this work are included. However, most of the used toolkits also allow for optional use of a seed translation dictionary, which tends to increase performance across the board.

The specifics behind alignment and projection are also dependent on the general topology of the embeddings that are to be mapped. In this work, all experiments are performed over Word2Vec embeddings, which offers two possible topologies: continuous bag-of-words (CBOW) and skip-gram. The former trains a shallow network to predict a word given an input context, while the latter learns to predict a context window from an input word. In this work, only skip-gram is used for all experiments, particularly skip-gram with negative sampling (SGNS) (Mikolov, Sutskever, et al., 2013). Therefore, from this point all references to the word embeddings used by the cross-lingual embeddings described can be assumed to be referring to this model. The training objective of the skip-gram model is to predict a context given a word, that is taken to be the center word the predicted context, as shown in figure 3.1. Because the softmax layer of the network corresponds with the generated context, its size its proportional to said context, which makes skip-gram more expensive to train when compared to CBOW for most non-trivial window sizes. However, skip-gram has been shown to outperform CBOW in capturing infrequent words effectively, therefore faring better in semantic tasks and comparably in syntactic ones (Mikolov, Chen, et al., 2013).

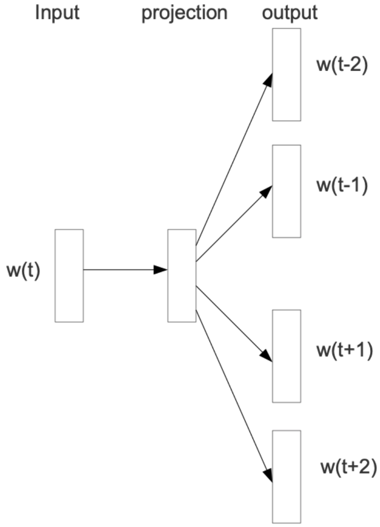


Figure 3.1: Neural network architecture for the skip-gram model (Mikolov, Sutskever, et al., 2013). Given a word w at training time, the network tries to predict its surrounding context, taking the input word as the center element of the predicted window context.

Projection-based CLE methods generally have a very similar operative, but differ from each other in how they accomplish each of the requirements for a common cross-lingual mapping. The general common steps can be described as follows, taking E_{L1} and E_{L2} to be monolingual word embeddings in two different languages:

1. Build a seed translation dictionary. Some models skip this step and instead require that a bilingual dictionary for the considered languages is given directly, in what is known as semi-supervision. Fully-unsupervised approaches induce this dictionary from the source monolingual vector spaces, assuming approximate isomorphism between E_{L1} and E_{L2}.

2. Use the seed translation dictionary to align monolingual subspaces E_{L1} and E_{L2}. This alignment is done over a hypothetical common cross-lingual space where certain dimensions with that are contain more significant information can be boosted. The system then iterates through each translation pair, and performs transformations over their corresponding vectors to situate them close to each other in a shared vector space.

3. Learn to project E_{L1} and E_{L2} to a shared cross-lingual space based on the previously aligned embeddings, maximizing mutual information.

3.1.1 Adversarial alignment: MUSE

MUSE (Lample, Denoyer, and Ranzato, 2017) uses adversarial training to create a generator network able to project word vectors from each of the given monolingual embeddings in a way such that it is very hard to distinguish the space to which they originally belonged – thus achieving a common mapping between both embeddings. This projection is then fine-tuned and the best model is chosen in the end, all without the need of any bilingual supervision. In greater detail, considering E_{L1} and E_{L2} to be monolingual word embeddings in two different languages and W_{L1} the projection matrix, the method is structured as follows:

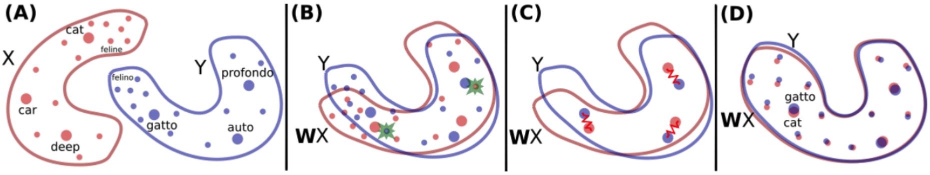


Figure 3.2: Diagram illustrating the operative of the MUSE model (Lample, Denoyer, and Ranzato, 2017). In it, the word vector space X is aligned with vector space Y using a proxy of the projection matrix W learned through adversarial training (step A). The matrix is then fine-tuned, using as guidance a set of induced translation pairs (step C). Finally, the best projection is chosen (step D).

1. **Adversarial training.** An initial proxy of the projection matrix W_{L1} is learned. In this step, W_{L1} corresponds to the generator component of the adversarial network, and it is defined as a linear map. The other element of the network, the discriminator, is a feed-forward network that tries to distinguish between true $L2$ vectors from E_{L2} and $L1$ vectors from the projection, $E_{L1}W_{L1}$. The joint objective function of this adversarial setup is then defined as a combination of the discriminator objective – which indicates how effective the discriminator is – and the mapping objective – that is, the objective of the generator –. The mapping objective improves as the generator is able to project word embeddings from both monolingual embeddings into an increasingly indistinguishable representation space. The joint learning algorithm follows the standard training procedure detailed in Goodfellow et al. (2014).

2. **Synthetic dictionary extraction.** A synthetic parallel vocabulary is built by considering the most frequent words and keeping only mutually-closest vectors, which increases the quality of the selected pairs. However, k-nearest neighbors (K-NN) present some problems when working with high-dimensional spaces, like

those defined by neural word embeddings. K-NN is an asymmetric criterion: if point y is a K-NN of x, it does not necessarily follow that x is a K-NN of y. In a high-dimensional representation space, this asymmetry leads to a phenomenon in which some vectors are likely to be the nearest neighbors of many other points, while others, are far removed from all data points (Radovanović, Nanopoulos, and Ivanović, 2010; Lazaridou, Dinu, and Baroni, 2015). To alleviate this issue, the authors propose cross-domain similarity local scaling (CSLS) as a distance metric. CSLS is defined as two times the cosine distance between a projected word vector E_{L1_t} from E_{L1} to another vector E_{L2_s} in E_{L2}, minus the mean cosine similarity of said projected vector to its target neighborhood in E_{L2} and minus the mean cosine similarity of the target vector from E_{L2} to the source embedding neighborhood in E_{L1}. It can be expressed as:

$$CSLS(E_{L1_t}, E_{L2_s}) = 2\cos(E_{L1_t}, E_{L2_s}) - r_T(E_{L1_s}) - r_S(E_{L2_t}) \tag{3.1}$$

Where function cos() corresponds to the cosine similarity calculation, $r_T()$ to the mean cosine similarity of the parameter vector and its target embedding neighborhood in E_{L2} and $r_S()$ is the mean cosine similarity of the parameter vector and source embedding neighborhood in E_{L1}.

3. **Mapping fine-tuning.** Improve the GAN-induced mapping W_{L1} through a refinement step based on a bootstrapping extension of the Procrustes problem as proposed in Schönemann (1966). Only a very small number of iterations – usually up to 5 – are needed.

4. **Unsupervised selection of the best model.** Use CSLS to generate a translation for each of the 10k most frequent words, then compute the average cosine similarity between the translations. This average is taken as an estimation of the performance of the model, meaning that the model with the highest average is considered the best one, and therefore it is selected. As a result, the metric can also be used as a stopping criterion.

3.1.2 Monolingual similarity distribution assumption: VecMap

Artetxe, Labaka, and Agirre (2018) assume that word translations have approximately identical vectors of monolingual similarity distribution. The proposed method operates on top of this assumption, adding empirically motivated enhancements that makes the procedure more robust. The overall process can be broken down into four steps:

1. **Embedding normalization.** VecMap uses multi-step pre-processing: first, it normalizes the embeddings with unit length normalization, then it applies mean centering, and finally unit length normalization again. The last normalization ensures that the resulting embeddings present unit length. Consequentially, the dot product of any pair of pre-processed embeddings is equivalent to their cosine similarity, which is a common similarity measure for word vectors. Additionally, it is directly related to their Euclidean distance. As an extra pre-processing,

zero-phase whitening filters (ZCA), also known as ZCA whitening (Bell and Sejnowski, 1997), is applied to both embeddings. The whitening encourages the exploration of dimensions that may not fit the current solution during the self-learning procedure, as a measure to escape poor local optima.

2. **Fully unsupervised initialization.** The seed bilingual dictionary is equal to the set of nearest neighbors according to the similarity between monolingual similarity distribution vectors. It is assumed that the embedding spaces are perfectly isometric, and therefore both axes of their respective similarity matrices correspond to words, which can be exploited to reduce the mismatch to a single axis. That is, the similarity matrices would be equivalent up to a permutation of their rows and columns, where the permutation in question defines the dictionary across both languages. In practice, embeddings may not always present isometry, but the property is treated as though it always holds approximately.

3. **Robust self-learning.** The method reaches the desired cross-lingual mapping through a self-learning bootstrapping procedure based on the Procrustes problem, similar to that of MUSE. Each iteration contains two steps: orthogonal mapping optimization over the seed bilingual dictionary and computing the optimal dictionary over the similarity matrix of the mapped embeddings. However, the procedure tends to get stuck in poor local optima when the quality seed translation dictionary is not good enough. As a result, some modifications should be applied over the seed dictionary. 1) Stochastic dictionary induction is performed: elements of the similarity matrix are randomly set to zero, with varying probability across iterations. 2) Before the self-learning step is started, the dictionary induction process is restricted to the k most frequent words in each language, where k is commonly equal to 20,000. 3) CSLS-based language pair retrieval is used instead of nearest neighbors in the second step of each iteration. 4) The dictionary is induced in both directions in order to avoid that repeated target language words trap the local optima.

4. **Symmetric re-weighting.** Once self-learning has converged to an acceptable solution, cross-correlational re-weighting is applied by de-whitening the embeddings that were initially whitened with ZCA filters. This boosts the weight of dimensions that best match the obtained solution.

3.2 Concatenation of monolingual corpora

Another approach to cross-lingual embeddings explored in this work are embeddings trained over multilingual corpora. Pseudo-cross-lingual embeddings use explicit or implicit cross-linguality and integrate it in the training corpora. For instance, in Gouws and Søgaard (2015) the authors concatenate the source and target corpus and replace each word that is part of a translation pair – obtained using a machine translation system – with its translation equivalent with a probability of 50%. Duong et al. (2016) expand on this by replacing each word center of the context with its translation at the same time that the system is being trained training. Semantic clusters have also been used to induce a weak bilingual signal (Ammar, Mulcaire, Ballesteros, et al., 2016), as well as shuffling document-level aligned data (Vulic and Moens, 2015).

However, the model proposed in this section is trained over corpora with no alignment or contextual proximity. The procedure is remarkably simple: two monolingual corpora in different languages are concatenated, and a word embedding is trained over the resulting multilingual text. This approach is actually used by some neural language models such as Lample and Conneau (2019) who, instead of inferring words in the source and target language sentences via skip-gram, predict randomly masked words in both sentences with a neural language model, in what is known as a masked language model (Devlin et al., 2019). Their model can be trained with or without parallel data, the latter working over concatenated monolingual datasets.

In the case of this work, this approach serves as a baseline to ascertain how much of the alignment achieved from cross-lingual embeddings is innate to the distribution of both languages and can be extracted with no mapping procedures.The word embedding architecture used to this end is skip-gram with negative sampling, described in section 3.1.

3.3 Evaluation strategies

Cross-lingual mappings are commonly evaluated according to their bilingual lexicon induction (BLI) performance (Artetxe, Labaka, and Agirre, 2018; Lample, Denoyer, and Ranzato, 2017), which is the task of inducing word translations from monolingual corpora, using a bilingual dictionary as ground-truth. However, BLI is an intrinsic task and in many cases might not be indicative of the performance of the system in its actual downstream task, such as document classification, information retrieval or NMT pre-training, among others (Ammar, Mulcaire, Tsvetkov, et al., 2016; Bakarov, Suvorov, and Sochenkov, 2018; Glavaš et al., 2019). Moreover, BLI is rarely the final objective when using cross-lingual mappings. Instead, the generated cross-lingual representation space is used for language transfer in cross-lingual tasks (Ruder, 2017), where integrating an machine translation setup might be too costly and not yield a significant improvement. Therefore, it is desirable to provide a downstream evaluation in addition to a BLI assessment when publishing on a benchmark for a cross-lingual method.

Although this would be ideal, standards for which tasks and models should be tested for in downstream experiments are not widespread, although efforts have been made in this regard (Glavaš et al., 2019). This is likely due to the added complexity and processing power that comes with integrating diverse NLP tasks with a cross-mapping method. Some authors acknowledge the advantages of experimenting over downstream tasks, but may not actually be able to use this sort evaluation due to time constraints and the aforementioned lack of consensus (Vulic, Glavas, et al., 2019). This work has opted out of providing diverse downstream tasks evaluation as a result of constraints on time and resources, relying instead purely on BLI for cross-mapping evaluation. However, since the generated will also be evaluated pre-training for an

NMT model – a process that will be described in chapter 4 –, it is arguable that the downstream task for which the embeddings are built in this work will be taken into consideration. Even then, it is important to ask the reader that BLI evaluation should not be taken at face value as a complete measure of the effectiveness of a cross-mapping method.

Bilingual lexicon induction is usually evaluated as mean accuracy over a set of translation pairs. There exist several options when it comes to measuring the distance between the words in the source and target language in a common word vector space. Some of the more widely used ones are:

- **Euclidean nearest neighbors**. In Euclidean K-NN, the k closest points in space to a given word vector are retrieved and sorted according to their Euclidean distance to said word vector. Euclidean distance is calculated as the absolute value of the numerical difference of their coordinates. In a hyperdimensional space, the Euclidean distance might become very large for vectors that are anything less than very close together, making it very hard to distinguish non-extreme ranges, and contributing to creating areas with many close points and others with isolated vectors, a phenomenon known as "hubness" (Radovanović, Nanopoulos, and Ivanović, 2010; Lazaridou, Dinu, and Baroni, 2015).

- **Nearest neighbors based on cosine similarity**. Calculates the k closest points in space to a given word vector using cosine similarity. Cosine similarity represents the similarity between two non-zero vectors in an inner product space. It is equal to the cosine of the angle between these two vectors, and therefore it is also equivalent to their inner product when the vectors are normalized. This calculation based on the angle of the vectors is more robust in hyperdimensional spaces than the Euclidean distance, which makes this metric preferable when working with embeddings. However, K-NN is by definition an asymmetric measure, and the main culprit behind hubness (Radovanović, Nanopoulos, and Ivanović, 2010). For this reason, cosine similarity has also been used in combination with other methods in the field of cross-lingual word representations.

- **Softmax over cosine similarity**. Artetxe, Labaka, and Agirre (2019) use a softmax function over cosine similarity to match language pairs. The translation candidate for a given source language word is obtained by taking the $n = 100$ Euclidean nearest-neighbors in the target language and scoring them with a softmax function, whose temperature is calculated using maximum likelihood estimation (MLE) over a dictionary induced in the reverse direction – that is, target to source –. This transforms cosine similarity scores into action probabilities, which are interpreted as translation probabilities in this context. Considering the source language word vector E_{L1_t} and the target language word vector E_{L2_s}, the metric can be defined as it follows:

$$\phi(E_{L2_s}|E_{L1_t}) = \frac{\exp(\cos(E_{L1_t}, E_{L2_s})/\tau)}{\sum_{i=1}^{w(E_{L2})} \exp(\cos(E_{L1_t}, E_{L2_i})/\tau)} \tag{3.2}$$

Where $\phi(E_{L2_s}|E_{L1_t})$ is the probability of finding E_{L2_s} as the translation of E_{L1_t}, $w(E_{L2})$ is the number of word vectors that can be found in the target language embedding and cos() is the cosine similarity function. The temperature τ is estimated using MLE based on an induced target-source language translation dictionary.

- **Cross-domain similarity local scaling (CSLS)**. Although the previously described method of using softmax in combination with cosine similarity succeeds at abandoning K-NN to alleviate dimensionality artifacts, it also requires a more costly calculation, especially when dealing with a large vocabulary. Lample, Denoyer, and Ranzato (2017) propose CSLS, which obtains the difference between two times the cosine distance from the source to the target vector and the mean cosine similarities of the source and target word vectors to the neighborhood of the opposite representation space, as detailed in equation 3.1 within section 3.1.1. This smooths the distance calculation for isolate data points or those that belong to a dense cluster, as it takes the surrounding neighborhood in the opposite representation space in consideration.

For this work, CSLS has been chosen due to its low calculation costs, which is relevant when a large number of evaluations are needed, and its integration with the cross-lingual mapping toolkits that implement the VecMap[1] and MUSE[2] methods.

3.4 Influence of language similarity

Fully unsupervised cross-lingual mappings generally assume approximate isomorphism between between vector spaces when it comes to inducing the initial seed dictionary (Barone, 2016; Søgaard, Ruder, and Vulić, 2018). That is, it is presumed that the shapes of the geometries defined by the vectors in both representation spaces are approximately similar. However, this property is exclusively satisfied in monolingual word vectors trained for similar languages and domains due to the Zipfian phenomena in language (Zipf, 1950). As a result, when using fully unsupervised cross-mapping methods with very distant languages the projection learned tends to be deficient or completely unsuccessful (Glavaš et al., 2019).

This effect is very much present in the cross-mapping approaches used in this work. Fully unsupervised MUSE learns an initial proxy of the projection matrix through adversarial training, and uses said projection to generate a bilingual dictionary by retrieving closest vectors based on their CSLS distance. However, for the closest translation pairs to be correct, the isomorphism assumption needs to be satisfied, which leads to noticeably worse results for distantly related languages (Søgaard, Ruder, and Vulić, 2018). VecMap has shown to be more robust when working with languages that are not similar (Glavaš et al., 2019), which can be attributed to some of its empirically motivated processing steps that allow the method to focus on the most promising dimensions, such as unit length normalization, mean centering, and ZCA

[1] https://github.com/artetxem/vecmap

[2] https://github.com/facebookresearch/MUSE

whitening, as well as cross-correlational re-weighting. VecMap also relies on stochastic dictionary induction, which sets elements of the similarity matrix to 0 randomly, helping the learned projection to escape local optima. Still, there is a clear correlation between language similarity and performance because, even with optimizations that improve the robustness of the method, the procedure is based on the assumption that inherent similarities between the word vector spaces exist.

Although the effect of language similarity in cross-lingual mappings has been studied by many authors, there exists a pervasive issue in this area of study – and it extends to the natural language processing field as a whole – where language similarity tends to be taken to be the same as phylogenetic relatedness. Take as examples Schönemann (1966) and Vulic, Glavas, et al. (2019), which are otherwise great publications that have helped this work greatly. The assumption that languages that belong to the same family are similar does not always hold. Consider the monstrously large Indo-European family of languages, which contains anything from Hindi to Swedish, two languages that have very little in common. On the other hand, the Uralic family, separate from Indo-European, features Finnish, a language to which Swedish is undoubtedly more similar than Hindi due geographic proximity, even if both languages are still remarkably different and do not share a common root. If genetic ancestry is the sole determiner for similarity, this would also mean that we should give up on finding acceptable cross-lingual alignments with isolate languages such as Basque, even though this particular language has been used in neural machine translation models to a moderate degree of success given the scarcity of corpora (Jauregi Unanue et al., 2018).

It is therefore important to separate the concept of phylogenetic relatedness and typological similarity. The former is the subject of study of historical linguistics, which is mostly interested in tracking diachronic language change to ascertain the origin and evolution of languages. The latter is focused on a synchronic survey of the features that appear in languages, and it can be directly equated with the concept of language similarity. Although languages that are close to each other in phylogenetic terms tend to also be typologically similar, this does not always happens, as the languages may diverge if they stay isolated from each other. In contrast, often times very genetically distant language may actually be quite similar due to contact thanks to geographic proximity. Part of the confusion comes from trying to interpret the classification of languages in the same way as that of living beings, a categorization system known as cladistics, which determines the similarity of two organisms using their last common ancestor[3]. After all we speak of phylogenetic relatedness between languages, a concept directly imported from biology. But this idea of similarity cannot be applied to languages, which can change considerably by interacting with each other as populations of speakers interact with other communities. Instead, it is imperative to base any similarity groupings on the typological characteristics found in languages, as clustering based on feature similarity has shown comparable or greater levels typological similarity than genetic grouping (Georgi, Xia, and W. Lewis, 2010).

[3] For example, zebra fishes are more closely related to an actual zebra than to sharks, since both of the striped animals share a lobbed-finned fish as a common ancestor, but the closest common ancestor of sharks and zebra fish is the first bony vertebrate.

In the case of this work, four language pairs are considered, which are in decreasing order of genetic relatedness: German-English, French-English, Russian-English and Hindi-English. All of them are included in the Indo-European family of languages and it is hypothesized that they share a common root. However, this grouping is extensive enough to include very different languages, as those from the Indo-Iranian subfamily such as Hindi differ significantly from languages belonging to the European subfamily – which includes English, German, French and Russian –, allowing for comparison between distant languages. This selection has also been motivated by the interesting properties of the relationship triangle formed English, German and French. English and German are the closest genetic relatives, and both are included in the West Germanic family. Within this group, modern German belongs to the High German branch, while English derives from the Anglo-Frisian subfamily. On the other hand, French belongs to the Romance family, which shares the grouping of European language along with all Germanic languages. This means that German and English are similarly distant to French according to their phylogenetic classification. German and French are indeed quite different. However, French and English share many more typological features, notably an estimated 25% of English loanwords come from French (Cannon, 1989), along with other characteristics heavily influenced by the French language (Pyles and Algeo, 1993). This three-way relationship is interesting because it juxtaposes genetic and typological features, both of which can have different effects over cross-lingual mappings. Aditionally, Russian and Hindi provide increasingly more distant languages that help study projection methods for cases with low cross-lingual similarity and different alphabets.

Chapter 4

Cross-lingual pre-training in neural machine translation

4.1 Cross-lingual word embedding pre-training

Word embeddings have been widely employed in multitude of NLP tasks, as they provide a robust semantic encoding that can be useful in any number of problems that operate with text, from document classification to information retrieval or text understanding (Naseem et al., 2020). This strategy to encode a lexical unit as a vector before it is used by a model has also proved to be valuable when applied to neural machine translation. The neural network learns its own contextual encoding for each word in the vocabulary during training, which is no surprise, since they follow a very similar operative to that of neural embeddings such as Word2Vec (Mikolov, Chen, et al., 2013). However, NMT systems tend to suffer in low-resource scenarios where sufficiently large parallel corpora is unavailable, which is especially problematic if the encoding and decoding pieces of the network are randomly initialized and need to be fully learned during training. Pre-trained word embeddings take advantage of monolingual corpora, which is otherwise not very useful when it comes to training a supervised neural model, and provide a strong initialization that generally improves the performance of the model, especially in scenarios that present paucity of data (Qi et al., 2018). However, word embeddings are relatively small models that only encode information at the semantic level (Mikolov, Chen, et al., 2013; Pennington, Socher, and C. Manning, 2014). This means that they need to be applied over an acceptable amount of training corpora. That is, even excellent pre-training is limited by its role on the overall model, which needs to learn other core features of the language such as word ordering and multitude of grammatical dependencies.

The utility of word embeddings in machine translation pre-training is a seemingly promising precedent to apply cross-lingual embeddings in this area. It follows that these methods are well fit for the task, since they create a common representation space that is then shared between encoder and decoder, aiding the neural model in creating a hidden transformation from source to target language. However, prior results show very little correlation between embedding cross-mapping – be it supervised or unsupervised – and improvements in the performance of the translation neural network (Qi et al., 2018). Consequently, the use of cross-lingual methods is virtually non-existent for pre-training, while monolingual word embeddings keep being relatively common.

4.2 Cross-lingual language model pre-training

Neural language models are, as of the time of this work, the most ubiquitous pre-training strategy for deep neural networks in machine translation and many other NLP tasks. Often derived from some of the most successful early architectures such as GPT (Radford and Narasimhan, 2018) or BERT (Devlin et al., 2019) – the latter being perhaps the most widespread approach of all –, neural language models are large neural networks – as opposed to the traditionally shallow models of word embeddings – that are trained over monolingual data and learn an internal representation of the elements of the language at hand. Their objective is predicting the words for the same language that they are trained over, for which they take into consideration a larger amount of contextual information than the embeddings, which rely only on semantic characteristics (Liu et al., 2020). The learned representation is hidden, meaning that it cannot be used as a general encoding, but instead the resulting language models should be integrated with another neural network that is designed with the final downstream task in mind. In the case of state-of-the-art machine translation, this secondary network tends to rely on some derivation of the Transformer architecture (Vaswani et al., 2017). As with word embeddings pre-training, neural language model pre-training consistently shows improvement over a randomly initialized baseline, with more prominent gains for low-resource language pairs where parallel data is scarce. The effect is more pronounced in the direction of the language for which there is more data (Liu et al., 2020), also a common phenomenon in word embedding pre-training.

Neural language models are large and have considerable number of parameters to estimate, which makes them more costly to train compared to word embeddings models. While being a much heavier model has obvious disadvantages, particularly in regards to the computing power needed to successfully train a neural language model, it also provides a stronger initialization. Once the task-specific network is integrated, it only needs to be fine-tuned over supervised data, since the neural language model already has learned a very complete representation of the input language (M. Lewis et al., 2020).

The concept of semi-supervised and fully-unsupervised cross-linguality has also been explored in neural language models. Lample and Conneau (2019) propose different approaches that rely on causal language modeling (CLM), masked language modeling (MLM) – derived from the MLM objective in Devlin et al. (2019), also known as the Cloze task (Taylor, 1953) – and translation language modeling (TLM) combined with MLM. The CLM and MLM objectives can be trained over monolingual data, while TLM requires parallel corpora. The cross-lingual language model is trained over monolingual corpora in both the source and target language, sampled in single-language batches. The authors apply byte-pair encoding (BPE) (Sennrich, Haddow, and Birch, 2015) to the training data and use a joint vocabulary for both languages, which encourages learning a shared mapping.

Also iterating on the aforementioned CLM objective, Ren et al. (2019) propose a cross-lingual masked language model (CMLM), that randomly chooses n-grams from the input text stream and predicts their translation candidates for each time step. Rather than relying mostly on BPE to learn subword information as in the case of Lample and Conneau (2019), the authors work over n-gram word embeddings (Bojanowski et al., 2016) learned on the monolingual corpus. Moreover, they incorporate an explicit cross-lingual signal, as the objective in training is to predict the translation of the input n-grams, similarly to the CLM + TLM setup in Lample and Conneau (2019). However, n-gram translation table is not hand-crafted, but instead unsupervisedly induced by applying the VecMap cross-mapping procedure (Artetxe, Labaka, and Agirre, 2018), and taking the translation probabilities as the similarity scores obtained using the marginal-based scoring method (Conneau, Lample, Ranzato, et al., 2017; Artetxe and Schwenk, 2019). As with the proposal of Lample and Conneau (2019), the effects on the translation of language pairs are statistically significant and overall more successful than previous attempts at NMT pre-training with cross-lingual word embeddings.

Given that language models have proved to be more effective with unsupervised cross-lingual approaches than word embeddings when it comes to pre-training neural models, they were initially considered for the experiments that appear in this work. However, they present some problems that have ultimately made their inclusion impossible. First and foremost, they are much more costly to train than neural word embeddings. This an especially harsh obstacle for this work, which intends to word with four language pairs whilst evaluating the impact of a variety of factors – namely the cross-mapping method used and whether BPE tokenization is applied –. As a result, the total number of models needed is considerable, making the whole setup very sensible to the individual training times of whichever methods are studied. Additionally, neural language models learn a hidden representation, which makes evaluating their performance as standalone models harder, requiring more complex tasks over natural language inference corpora such as Conneau, Lample, Rinott, et al. (2018), which are less suited for word embeddings. This difficulty in comparing both pre-training models is to be expected, since they are not really designed to do the same things – word embeddings learn purely semantic information while neural word embeddings are able to learn more diverse contextual information –. Since the priority in this work is to ascertain the effect of certain variables when working with cross-lingual word representation and their effect on pre-training rather than producing a top-of-the-line system, and computational resources are limited, it has been decided to drop the experimentation from neural language models. They, however, remain an extremely interesting field of study that has been rapidly growing and showcasing consistent improvement.

4.3 Pre-training limitations

As mentioned previously, the effect of cross-linguality in pre-training word embeddings has been disappointing in many cases (Qi et al., 2018). When it comes to the use of non-aligned monolingual vectors, the improvements in performance from models without and with pre-training seem to be larger the more similar that the source and target languages. This, along with the low impact that cross-lingual methods have been shown to have on pre-training, seems to indicate that the neural network is learning its own projection from source to target language along with whichever other transformation fits its training objective – i.e. translation or a discriminant function –. This behavior has been understood for some time now, especially in the case of Transformer-derived architectures (Vaswani et al., 2017), which have specific sections of the network dedicated to projecting to and from incoming text into a specific representation, known the encoder and the decoder modules. However, the neural network itself is also limited when it comes to generating a good projection if not enough parallel data is given, and said projection may also be harder to learn depending on the degree of similarity of the language pair at hand (Lample and Conneau, 2019). Therefore, cross-lingual pre-training is still desirable in order to take full advantage of monolingual corpora to initialize the network. Cross-lingual neural language models have been more successful at improving prior benchmarks (Lample and Conneau, 2019; Ren et al., 2019), but the correlation between their performance in evaluations of their cross-lingual representation (Conneau, Lample, Rinott, et al., 2018) and that of the final fine-tuned network is still relatively weak. Sun et al. (2019) investigate the interaction between unsupervised cross-lingual embeddings and unsupervised NMT, and point to a phenomenon that may explain this seemingly common problem to pre-training methods in charge of learning a representation of the language. The authors argue that, during the fine-tuning phase, the neural network may refine the projection provided by the pre-trained layers, taking advantage of the fact that it has already been initialized. But since the model is training towards an independent objective, it may heavily modify the initial representation in order to fit its optimization criteria, leading to degeneration of the pre-trained embeddings or neural language model. To alleviate this issue, Sun et al. (2019) propose to train the NMT mode objective – in this case translation, expressed as agreement regularization – jointly another one that aims to maximize the correspondence of the encoder representation space – corresponding to the source language embedding – and that of the decoder, represented by the target language embedding. This latter objective is implemented through adversarial training, where a discriminator is trained to maximize the probability of choosing the accurate language to which the word vector belongs, that is, if it belongs to the transformation applied by the projection matrix in conjunction with the encoder or if it comes from the decoder space. A generator that takes the cross-lingual projection matrix as a trainable parameter tries to confuse the aforementioned discriminator in what is a two-player minimax game (Goodfellow et al., 2014). Results seem to indicate that the degeneration of the pre-trained spaces is at least partially responsible for previous experiments where cross-linguality seemed to have no effect, and training with a joint objective is desirable. Most recently, Wang and

Zhao (2021) build on the premise of Sun et al. (2019) and apply joint optimization to cross-lingual neural language models integrated in both supervised and unsupervised NMT models, as well as adding denoising to the model, with promising results.

Chapter 5

Experimental setup

5.1 Corpora

The corpora used in this work were originally selected from the data collection provided in the WMT14 Machine Translation shared task[1] (Macháček and Bojar, 2014). This choice was made to facilitate comparisons with other translation systems, since most of the datasets are well known and their use is widespread in the field of natural language processing. As a result, both the models submitted to the shared task and the many that use a similar combination of corpora sources can be used as reference for performance. The corpora also contain a considerable volume of monolingual data, which is vital for the unsupervised pre-training techniques explored in this work.

This collection of corpora also allows to study language similarity as a variable, since all of the language pairs available feature English data aligned with languages with which it presents varying degrees of phylogenetic relatedness and typological similarity. As detailed in section 3.4, the language pairs chosen to evaluate this effect have been French-English, German-English, Russian-English and Hindi-English.

The monolingual training set is a combination of some of the corpora provided in the WMT14 Machine Translation shared task, particularly the News Crawl (provided by the task) and HindMonoCorp (Bojar et al., 2014) collections. The latter has been added due to the very small size of the Hindi section of the News Crawl, especially in comparison with the other languages. Keeping a roughly similar volume of training data between for monolingual embeddings that are to be cross-mapped into a bilingual space is usually desirable, since most cross-mapping methods rely, partially or totally, on some sort of inherent shared similarity between languages. For this reason, the full News Crawl corpus is not used for the English and German monolingual corpora, reducing their size to approximate that of the other languages.

Similarly, the parallel corpora used to train the NMT Transformer model has been built by combining different corpora in such a way that the volume of data is comparable across all languages.

[1] https://www.statmt.org/wmt14/translation-task.html

Language	Corpora	Sentences	Tokens
English	News Crawl 2011-2013	51M	1,167M
French	News Crawl 2007-2013	30M	696M
German	News Crawl 2012-2013	55M	970M
Russian	News Crawl 2007-2013	32M	576M
Hindi	News Crawl 2007-2013 + HindMonoCorp 0.5	43M	932M

Table 5.1: Breakdown of training monolingual corpora sources, number of sentences and total number of tokens by source.

Language pair	Corpora	Sentences	Tokens	
			Source	English
French-English	Europarlv7 + Common Crawl corpus	5M	117M	129M
German-English	Europarlv7 + Common Crawl corpus	4.1M	99M	94M
Russian-English	Common Crawl corpus + Yandex 1M corpus v1.3	1.8M	41M	39.5M
Hindi-English	HindiEnCorp 0.5	0.26M	2.6M	4.1M

Table 5.2: Breakdown of training parallel corpora sources, number of sentences and total number of tokens by source.

The corpora sets used have the following common pre-processing steps:

1. Normalization of unicode punctuation encoding.

2. Tokenization.

3. Clean and eliminate empty sentences, those containing more than 60 words and sentences with a source-target ratio greater than 1-9.

These operations have been performed using software available in the Moses toolkit (Koehn, Hoang, et al., 2007). All the described datasets can be compiled using the guide and code used in this word, which is publicly available[2].

In some experiments, Byte Pair Encoding, also known as BPE (Sennrich, Haddow, and Birch, 2015), is applied to the corpora in order to study its effect in the performance of cross-lingual embedding mappings. BPE was originally conceived as a basic data compression technique which operated by finding the most frequently occurring pairs of adjacent bytes in the data and replacing all instances of the pair with a byte that was not in the original data, repeating the process until no further compression is possible. Applied to text compression, this algorithm functions at the character level. However, an adaptation of this method to subword-level compression (Sennrich, Haddow, and Birch, 2015) has risen to prominence in the natural language processing field since its conception. This version of Byte Pair Encoding, which is the

[2] `https://github.com/MarTnquesada/tfm`

one used in this work and that will be referred to as simply "BPE" from this point onwards, provides features that are far more beneficial for natural language processing tasks – and especially for machine translation – than simple file size compression. BPE encodes the vocabulary of a given corpus into its most common co-occurrent subword units, which can be combined to produce any of the words seen in training. This process reduces the total size of size of the vocabulary, which speeds up machine translation models and reduces their memory requirements. Most importantly, the subword units can be combined to represent not only the words in the training vocabulary, but also those out of it, that is, words that do not appear in the training corpus. Therefore, it allows translation models to operate with a de facto open vocabulary. This work studies the influence of joint BPE encodings, which build a shared vocabulary for the source and target language corpora. As a result, some lexical information is transferred more effectively between languages, particularly in the case of certain word classes such as instance names, compounds, cognates and loanwords (Sennrich, Haddow, and Birch, 2015).

The BPE implementation used in this work is fastBPE[3], which ports the software provided in Sennrich, Haddow, and Birch (2015) to the programming language C++ with the aim to improve its performance.

5.2 Tools

To generate the embeddings used in this work, version 0.9.2 of the FastText[4] (Bojanowski et al., 2016) toolkit has been used. FastText is an implementation of the Word2Vec architecture that extends the CBOW and skip-gram models by learning subword information, although this feature has not been used to allow for a fair comparison between cross-mapping methods, since some of them are unable to take advantage of character-level data. In particular, the model used is skip-gram with negative sampling, which follows the Word2Vec specifications that have been described in section 3.1. In the case of the cross-mapping techniques applied in the experiments, both MUSE[5] (Lample, Denoyer, and Ranzato, 2017) and VecMap[6] (Artetxe, Labaka, and Agirre, 2018) have open source implementation that have been kept functional since their respective original publications, and which have been used in this work. Concerning MUSE, a few slight alterations have been made in order to eliminate the partial supervision present by default during the self-evaluation phase of the cross-mapping procedure, leaving it as a fully unsupervised method. This modification has been created for the purposes of this work and is available as fork of the original MUSE project[7]. Finally, concatenation of monolingual corpora does not need of any additional tools – since there is no explicit projection, merely a FastText embedding trained on multilingual data –.

[3] https://github.com/glample/fastBPE

[4] https://fasttext.cc/

[5] https://github.com/facebookresearch/MUSE

[6] https://github.com/artetxem/vecmap

[7] https://github.com/MarTnquesada/MUSE

For the neural machine translation model used in combination with embedding pre-training, the NMT toolkit OpenNMT[8] (Klein et al., 2017) has been chosen. Specifically, version 2.1.2 of the OpenNMT PyTorch-based (Paszke et al., 2019) implementation dubbed OpenNMT-py[9] has been used. All launch configurations of the different models used in the experiments presented here can be found in the public repository of this work[10].

5.3 Evaluation metrics

5.3.1 Bilingual lexicon induction evaluation

The quality of the different cross-lingual representations generated is calculated as the average accuracy of said vector space when finding translation pairs. The evaluation considers the source language words from a bilingual ground-truth dictionary that are also included in the vocabulary of the source language embedding. For each of these source language words, the closest word vector in the target embedding is found and taken as the most likely translation. The procedure then compares the translations obtained with this method and those provided by the bilingual dictionary, taking as correct a the translation induced from the cross-lingual space if the target language closest vector is included in the list of possible translations that appears in the bilingual dictionary. The distance between word vectors is calculated using cross-domain similarity local scaling (CSLS), which has been been described in detail in section 3.1.1. This metric is designed alleviate some of the hubness (Radovanović, Nanopoulos, and Ivanović, 2010) phenomena in high-dimensional spaces. Finally, it obtains the average accuracy of the system. The implementation for this evaluation is a modification of the version proposed in VecMap designed for this work and available in the repository of this project[11], where BPE tokenization of the bilingual dictionary has been made available. The ground-truth bilingual dictionaries used are provided by the MUSE toolkit[12], specifically those belonging to the "full" set.

5.3.2 Machine translation evaluation

The neural OpenNMT machine translation models that use pre-trained embeddings are evaluated based on their translation accuracy using BLEU. Both the multi-BLEU criterion implemented by the Moses toolkit (Koehn, Hoang, et al., 2007) and the SacreBLEU library (Post, 2018) have been used, and their results are equivalent in all cases or differ in the order of single tenths of a point. Multi-BLEU has the disadvantage of requiring user-supplied pre-processing. However, since this work provides the instructions required to compile the corpora used in this work, it has been considered that the information provided is sufficient to reproduce the results of the experiments

[8] https://opennmt.net/

[9] https://github.com/OpenNMT/OpenNMT-py

[10] https://github.com/MarTnquesada/tfm

[11] https://github.com/MarTnquesada/tfm

[12] https://github.com/facebookresearch/MUSE

presented here using multi-BLEU if the reader so desires. SacreBLEU, on the other hand, expects detokenized outputs and applies pre-processing transformation on its own, providing a breakdown of the tokenization and pre-processing parameters used upon completion. Its integration with the WMT datasets makes SacreBLEU especially adequate for this task, although it should be highlighted that the test set used in the experiments of this work corresponds with that of the original WMT14 call[13] rather than the WMT14 corpus included in the SacreBLEU toolkit.

5.4 Model configuration

5.4.1 Cross-lingual models

5.4.1.1 Embedding configuration

All skip-gram word embeddings use in the experiments proposed in this work are trained during 5 epochs with a learning rate of 0.05. Changes in epoch size have not had any significant effect, as the corpora used to train the embeddings is sufficiently large and does not require of more training cycles. The main parameter to study when training the embeddings has been their dimension. Embeddings that use the Word2Vec architecture generally use a dimension parameter that tends to be between 50 and 300 (Mikolov, Chen, et al., 2013), although higher values are also sometimes considered. In work, a range of values between 100 and 1000, with steps of size 100, is used to examine the effect of embedding dimension in unsupervised cross-lingual methods. The results are displayed in figure 5.1 as an evolution of the BLI performance of the model relative to the dimensionality of the embeddings.

In the case of the cross-mapping strategies of VecMap and MUSE, the graphs seem to indicate that BLI increases rapidly with dimension up to close to 300 dimensions. From here to 500 dimensions, performance still seems to be correlated to dimension, but growth is very low. Up to 1000 dimensions, performance increases extremely slowly. This behaviour is in line with the directives originally given for Word2Vec embeddings, where very high dimensionality does not seem to provide a particular improvement in representation quality over the commonly used 300 dimension value (Mikolov, Chen, et al., 2013). Since any embedding with dimension equal or larger than 300 or 400 has peaked in terms of cross-lingual performance, choosing the value of this parameter will also have to do with the final system and task in which the vector space is used, and often extreme values such as those close to a 1000 will be avoided to avoid scarcity and hyperdimensionality-related problems. However, as seen in the section (d) of figure 5.1, relative to the hi-en language pair, it seems that increasing the dimension of the embeddings past a certain point may affect the viability of learning an effective cross-lingual projection for the adversarial approach in MUSE. In the absence of additional experiments with a larger number of distant languages, it is hypothesized that using too many dimensions while dealing with a complicated projection between very different languages that do not even share a common alphabet can lead to a failure in learning a reasonable projection matrix.

[13] `https://www.statmt.org/wmt14/translation-task.html`

VecMap is not affected by this phenomena for the showcased experiments, which may be due to the use of ZCA whitening (Bell and Sejnowski, 1997), which encourages exploring dimensions that may not fit the current solution to help escape poor local optima, a problem that is exacerbated when operating in a high-dimensional space. Additionally, VecMap also uses stochastic dictionary induction, where elements of the similarity matrix are set to zero at random, a technique that is also directed towards exploring a solution outside of the current optima.

For embeddings trained over concatenation of monolingual corpora, dimension does not affect their BLI score as a cross-lingual model. This is a testament of the effectiveness of CSLS as a measure of distance in high-dimensional spaces, since the overall representation is not changing as dimensions increase, but zones zones with extremely high or low density of data points can be generated for large dimension values. Concatenated corpora embeddings seem to fail to produce a meaningful alignment for language pairs with more complex projections such as ru-en or hi-en, and increasing the dimension does not seem to affect this phenomenon. Thus, embeddings learned over multilingual corpora do not appear to have their cross-lingual information affected if a dimension-invariant distance metric is used.

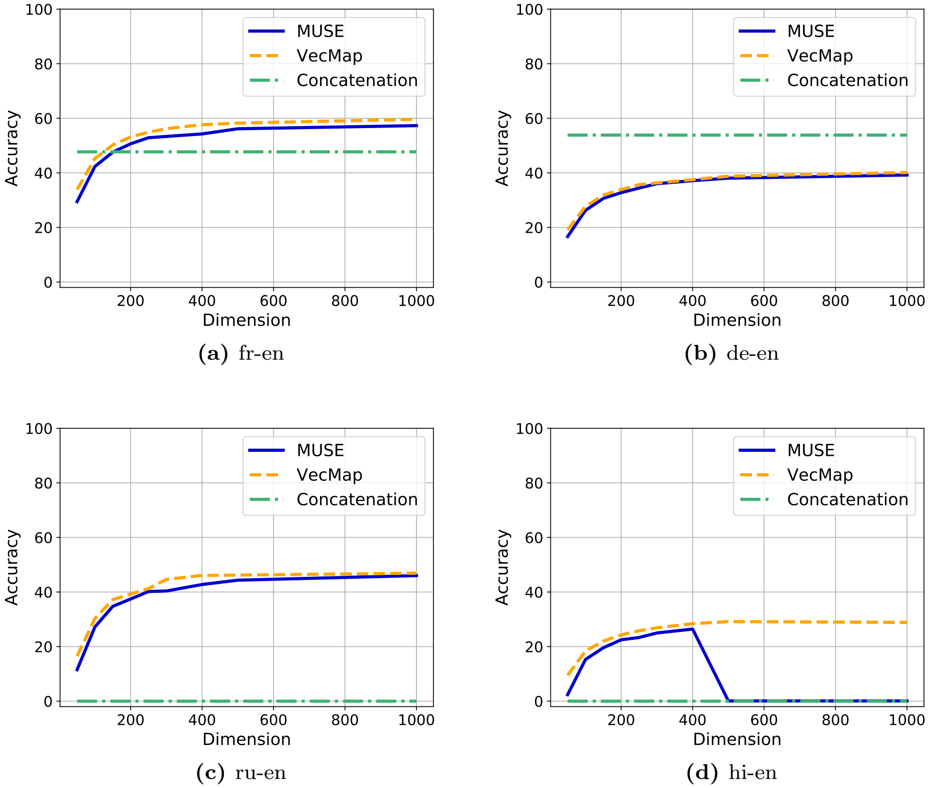
For this work, since the system in which the embeddings needs to be integrated is a Transformer neural network in charge of machine translation, a dimension value of 512 has been chosen as a fine compromise, since it is in the range where cross-lingual performance is stabilized, and corresponds for a common encoder-decoder dimension value for Transformer-derived machine translation models (Vaswani et al., 2017; Lample and Conneau, 2019).

5.4.1.2 Cross-mapping configuration

Both the MUSE and Vecmap cross-mapping techniques have a number of parameters that dictate some of the characteristics of the alignment procedure.

- VecMap uses the standard unsupervised configuration, which is equivalent to that of the models presented in Artetxe, Labaka, and Agirre (2018). The maximum vocabulary is set to 20,000 words, the vocabulary used to generate the initial unsupervised translation table is limited to 4,000 words, the CSLS neighborhood used for vector distance calculations is of size 10 and the embeddings are normalized before the cross-mapping is initiated.

- For MUSE, all alignments are performed using the default unsupervised parameters. The only explicit adjustments made are the maximum size of vocabulary considered, which is set to 20,000, and the number of word vectors used for discrimination, which is set to the 7,500 more frequent words. Distance between vectors is calculated using a CSLS neighborhood of size 10, and the embeddings are normalized before the cross-mapping process begins. The default unsupervised values caused memory problems in the setups available for this work, which meant that reproducing published VecMap configurations was far easier than that of MUSE. Therefore, the changes in this case are made to keep VecMap and MUSE running over with as similar of a set of parameters as possible.



(a) fr-en

(b) de-en

(c) ru-en

(d) hi-en

Figure 5.1: BLI performance evolution for each language pair and cross-mapping method according to the dimension of the monolingual embeddings. Each plot corresponds to a language pair, and the different series to one of the cross-mapping approaches considered.

5.4.2 Neural machine translation pre-training

The generated word embeddings are also assessed on their value as pre-training for a neural machine translation model. The architecture chosen is an OpenNMT-py Transformer model that mimics the original Transformer architecture proposed in (Vaswani et al., 2017). The model is an attention-based encoder-decoder network with 6-layered encoder and decoder, positional encoding, 8 attention heads and a dense feed-forward network between the two. The network uses a dropout probably of 0.1 both for the feed-forward layers and the attention heads, and relies on the Adam optimization criterion (Kingma and Ba, 2015). However, some aspects of the model have been adapted for this work:

- The feed-forward network that connects encoder and decoder has been changed to have a dimension of 1,024 units from the original 2,048 present in Vaswani et al. (2017). This accelerates model training and does not impact performance massively for models that are not trained extensively (Lample and Conneau, 2019).

- The number of training steps is reduced from 200,000 to 20,000, maintaining a batch size of 4,096 tokens from the original proposal. OpenNMT does not use epochs to determine training length, but instead steps, where each step is equivalent to training over a number of samples equivalent to the batch size, and samples are chosen at random. The decrease in training time is made due to time and resource constraints that would require the removal of some of the variables examined, since the priority of this work is to study the interactions of the different cross-mapping procedures when applied to a set of languages of varying typological features, rather than building a state-of-the-art model. Since pre-training can be very ineffective when the supervised translation model has not learned a minimum acceptable performance towards the objective that it is training towards (Qi et al., 2018), the final value of 20,000 steps has been chosen a compromise between training cost and minimum performance of the model to allow for pre-training integration.

Chapter 6

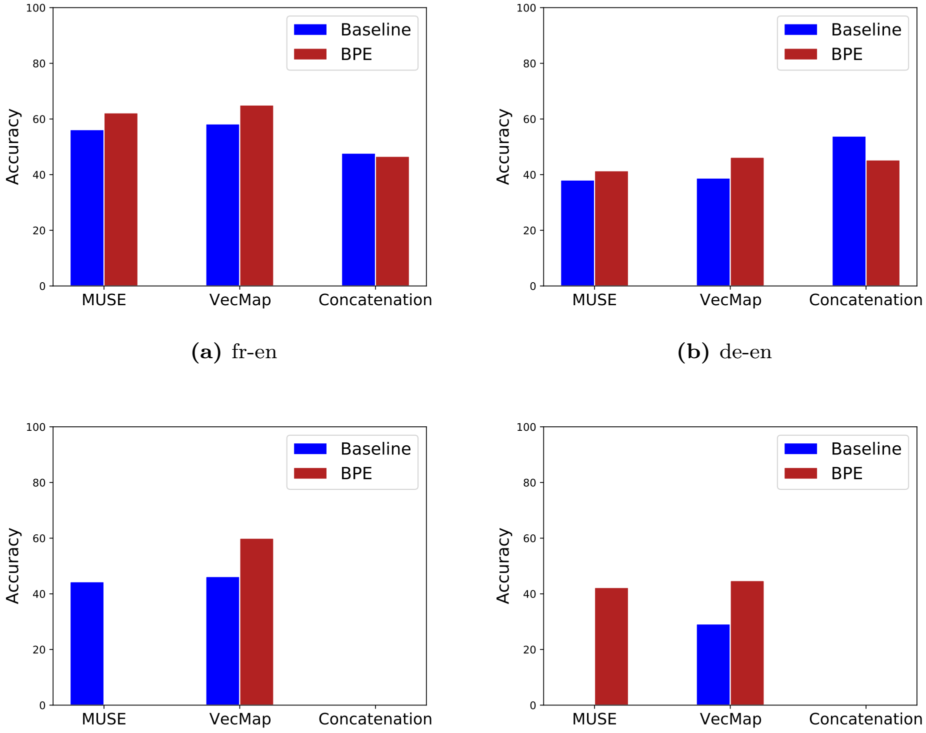
Results and discussion

6.1 Cross-lingual models

6.1.1 BPE

The cross-mapping method MUSE (Lample, Denoyer, and Ranzato, 2017) is able to take advantage of embeddings that include subword information such as FastText (Bojanowski et al., 2016), which has shown to have a positive effect on cross-linguality (Lample, Denoyer, and Ranzato, 2017). However, this feature has not been used in this work, since VecMap (Artetxe, Labaka, and Agirre, 2018) cannot capture character-level information, and therefore the inclusion of n-gram vectors would not allow for a comparable evaluation of the cross-mapping methods. To provide another type of subword information that can be used across the approaches considered, it has been decided to use BPE (Sennrich, Haddow, and Birch, 2015), which poses an opportunity to explore the effect of this widespread pre-processing technique in cross-lingual maps. In the past, Lample, Denoyer, and Ranzato (2017) have shown that BPE improves considerably the alignment of monolingual language spaces, particularly for cases where the languages share the same alphabet or anchor tokens (Smith et al., 2017). Anchor tokens are words with equivalent meaning that are written identically across languages, such as proper nouns of places, organizations or people – i.e. Berlin, Madrid, Google, Alan Turing –, acronyms – i.e. UE, UN, UPV, although some are variable depending on the language –, loanwords – i.e. siesta, croissant – and digits, which can perform the role of anchor tokens even between extremely different languages –.

Figure 6.1 illustrates the effect of BPE on the BLI performance of the generated cross-lingual embeddings. BPE usage seems to generally improve the score of the projection-based mappings, while having a slightly negative or non-existent influence on embeddings trained over concatenation of monolingual corpora. While the former result is expected (Lample, Denoyer, and Ranzato, 2017), the latter phenomenon is more interesting, and can be explained by the fact that these embeddings are jointly learning both languages, but no cross-mapping is performed, so the relative position of



(a) fr-en

(b) de-en

(c) ru-en

(d) hi-en

Figure 6.1: BLI performance comparison for each language pair and cross-mapping method between basic tokenization ("Baseline" in the figure) and additionally applying BPE tokenization ("BPE" in the figure). Refer to table 6.1 for detailed values.

words in the representation space should remain similar whether subword information is captured or not. Furthermore, in many cases there may be a certain loss of semantic information when creating a shared byte-paired encoding between languages (Ren et al., 2019), which often will be compensated by the subword features that are retrieved, but since this particular joint embedding approach does not take advantage of them, the overall effect of BPE tends to be negative.

The impact of BPE is especially significant for the MUSE mapping in language pairs ru-en and hi-en. In the case of ru-en, the use of this tokenization approach apparently does not allow for any sensible alignment, unlike the projection that uses non-BPE embeddings, which performs fine. A likely explanation for this is that Russian and English do not use the same alphabet, and therefore no joint subword information is learnt, while some semantic features may be diluted (Ren et al., 2019).

Additionally, the number of anchor tokens is very reduced, which further decreases the utility of BPE. However, the hi-en pair in figure Figure 6.1 shows the opposite phenomenon, where the application of BPE has made possible a previously unavailable alignment. This case is especially puzzling, since Hindi also uses a completely different alphabet from that of English, so even if is noticeably influenced by it, there should be very little transfer of information between the subword vocabularies. Upon closer inspection, both BPE vocabularies for ru-en and hi-en have an extremely similar size, which indicates that this behavior is not a function of semantic diversity. Instead, a possible explanation could be found that the generated vector spaces simply have a slightly different distribution when using BPE, which can affect the chances of finding a good alignment between embeddings, though more research is necessary to arrive at any definitive conclusion. The VecMap projection does not seem to be affected in the same way, which could be due to it having a more robust initialization and being able to escape local optima better than MUSE, as shown in previous publications (Vulic, Glavas, et al., 2019; Glavaš et al., 2019).

Although the effect of BPE has been generally benefitial in these experiments, Bostrom and Durrett (2020) show that unigram language model tokenization (Kudo, 2018) tends to outperform BPE in neural language model pre-training. In contrast to BPE, unigram language model tokenization initializes its base vocabulary to a large number of subwords and progressively trims them down to obtain a smaller vocabulary. The authors argue that unigram language model tokenization retrieves subword units that align more closely with morphology and avoids issues derived from the greedy construction procedure that BPE employs. Unigram language model tokenization could therefore also prove to be useful in cross-lingual embeddings, and may be interesting to explore in future work.

6.1.2 Language similarity

Table 6.1 showcases the performance of the considered cross-lingual models for all language pairs. Language similarity does seem to be somewhat indicative of BLI performance, although a weak signal at that. The fr-en language pair is the best performing one across the board, especially for the projection-based alignments. This is expected, since these methods are reliant on semantic similarities, and they make great use of anchor tokens in their initial unsupervised dictionary induction (Lample, Denoyer, and Ranzato, 2017), which should be plenty for this language pair given that English and French share a large number of words.

Although English and German are phylogenetically closer to each other, the performance for this pair is inferior to that of English and French for MUSE and VecMap. In contrast, embeddings trained over a concatenation of monolingual corpora surpass projection-based cross-maping methods, and their own BLI score for the fr-en pair. French and English share many anchor tokens, whereas German and English have a noticeably smaller common vocabulary, but show a greater degree of similarity in other typological features common in languages from the same family tree, such as word ordering or verbal categorization. As mentioned previously, projection-based cross-mapping approaches are highly dependent on cross-lingual similarities, so this behavior is easily explained. Since for the concatenation strategy the pair de-en is actually performing better than fr-en, it can be hypothesized that the natural alignment resulting of training embedding over multilingual text is more sensible to other typological categories. This is especially likely for word ordering, since the skip-gram architecture is learning to predict contexts in a reduced local window (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013), which is sensible to large discrepancies in sentence structure.

Conversely, training embeddings over a multilingual corpus seems to produce no alignment whatsoever when the selected languages do not share a common alphabet, such is the case for the ru-en and hi-en pairs. That said, it is not possible to assert this without more conclusive evidence, since languages with different alphabet that that present similar features in word ordering might fare better. For MUSE and VecMap, BLI performance seems to generally decrease with phylogenetic distance, as seen in prior work (Qi et al., 2018), though the effect is lesser than expected. In the particular case of MUSE, BPE presence has been shown to facilitate or invalidate the initial unsupervised mapping, which as discussed before is likely to be a function of distributional variation. For both of the projection-based cross-lingual techniques, ru-en and hi-en are shown to be surprisingly competitive with de-en. This casts some doubts on which are actually the typological features that govern explicit cross-linguality, since by all accounts German should be more semantically and grammatically similar to English than Russian or Hindi (Georgi, Xia, and W. Lewis, 2010). For example, German does not allow for noun clusters and instead builds agglutinates nouns into compounds, which might alter the horizon of the context window used when training the skip-gram embeddings. Compounding is also a common way to construct words for Russian and Hindi, though agglutinations tend to contain a smaller number of nouns and be significantly shorter. It might also be that Russian and Hindi contain a larger number of loanwords, which act as anchor tokens. Further research that isolates typological features for cross-lingual evaluation is needed in order to produce meaningful guidelines on the adaptation of cross-lingual models according to language similarity, though the experiments show that semantic relatedness is not the only factor at play.

Overall, VecMap is shown to be the best performing cross-lingual method and also the most robust one when dealing with distant languages, which lines up with previous research (Vulic, Glavas, et al., 2019; Glavaš et al., 2019). MUSE shows to be generally weaker for cases where inducing an initial translation table is more difficult due to low language similarity, though seems to perform fine when this phase is completed successfully, which is a common trend in projection-based cross-lingual methods. Surprisingly, training word embeddings over a concatenation of monolingual corpora outperforms projection-based methods for the de-en pair, although is not effective for distantly related languages. From this it can be inferred that some features learned by skip-gram word embeddings during training are valuable when it comes to producing an alignment, and, like many other typological characteristics, are not being considered by current explicit cross-mapping methods but could prove to be valuable.

Model	Dimension	BPE	Accuracy (cosine similarity)			
			fr-en	de-en	ru-en	hi-en
MUSE	512	No	56.2	38.1	44.3	0.1
MUSE	512	Yes	62.2	41.4	0.1	42.3
VecMap	512	No	58.2	38.8	46.2	29.2
VecMap	512	Yes	**65.0**	46.2	**60.0**	**44.8**
Concatenation	512	No	47.7	**53.8**	0	0
Concatenation	512	Yes	46.6	45.3	0	0

Table 6.1: Results obtained for the best cross-lingual embeddings selected for neural model pre-training. Accuracy is measured comparing pairs from ground-truth bilingual dictionaries and employing CSLS as distance metric, as described in section 5.3.1.

6.2 Neural machine translation pre-training

6.2.1 Freezing embeddings

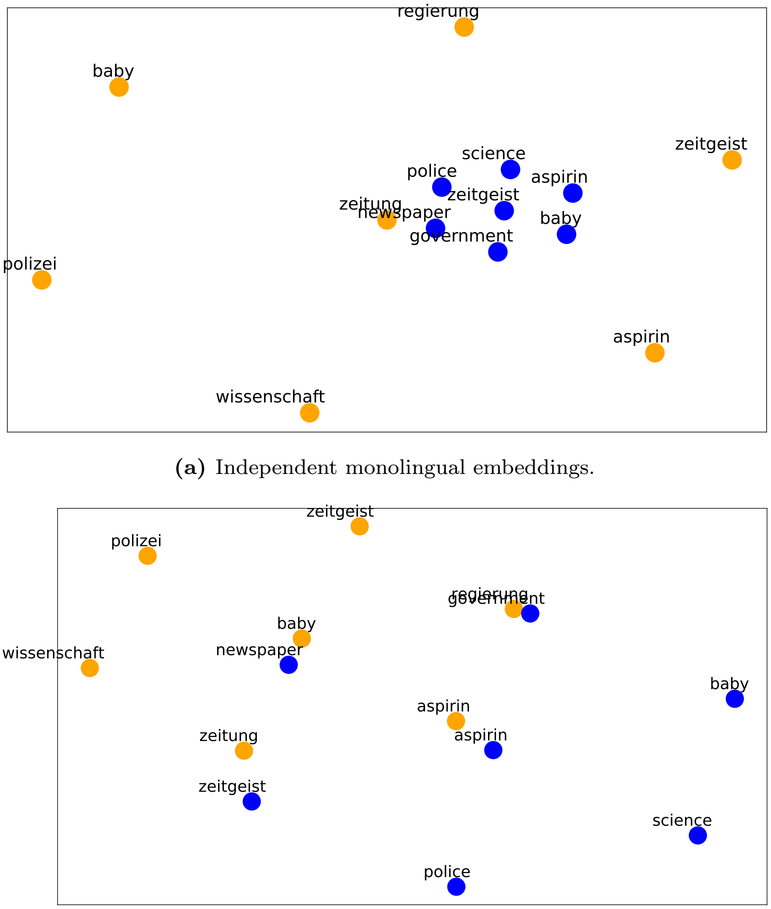
As indicated in Sun et al. (2019) and Wang and Zhao (2021), embeddings used as the encoder-decoder pieces of an attention-based neural network tend to degenerate as the global model is fine-tuned for a particular task, which for this work corresponds to machine translation. For this reason, it has been decided to assess the impact of freezing the encoder and decoder embeddings during training. The results are shown in table 6.2. Freezing the pre-trained embeddings does not improve BLEU performance in comparison with models that do modify the weights of their encoder and decoder during supervised training, which score slightly better. This effect is in line with prior work (Sun et al., 2019; Wang and Zhao, 2021), which shows that the integrated model needs to modify the pre-trained components during fine tuning to maximize its performance, but this behavior tends to break the cross-lingual alignment created previously. As a result, they propose to optimize supervised training based on two different objectives: maintaining the structural correspondence of the initial pre-trained components and maximizing the translation objective. Though implementations of this strategy are not widespread and therefore not readily available for general use – which is the reason why they have not been considered for these experiments –, they have been shown to be the best current approach to transfer cross-lingual knowledge in pre-training.

Pre-trained emb.			Frozen emb.	BLEU			
Cross-mapping	Dimension	BPE		fr-en	de-en	ru-en	hi-en
(None)	512	Yes	No	**34.1**	25.4	28.7	6.1
(None)	512	Yes	Yes	32.1	23.6	26.8	5.8
MUSE	512	Yes	No	33.9	26.0	29.4	6.7
MUSE	512	Yes	Yes	32.2	24.3	27.9	6.1
VecMap	512	Yes	No	33.4	**26.3**	30.0	**9.2**
VecMap	512	Yes	Yes	32.4	24.1	28.9	8.4
Concatenation	512	Yes	No	33.5	25.5	**30.2**	6.6
Concatenation	512	Yes	Yes	33.1	24.9	29.6	6.2

Table 6.2: Results obtained for the best Transformer translation models that use pre-trained vectors. Cross-mapping is indicated as (None) when no explicit cross-lingual technique is applied, that is, only monolingual word embeddings are used as encoder-decoder in the neural network.

6.2.2 Cross-linguality

The effect of cross-linguality is relatively low across the board, with the baseline setup that uses monolingual embeddings with no alignment for the encoder and decoder – indicated as (None) – outperforming the methods with explicit cross-lingual information in many language pairs. This behavior has been noted by previous authors (Qi et al., 2018), and can be partially attributed to the degeneration phenomenon (Sun et al., 2019) discussed in the previous section. However, it is also important to consider that the alignment generated by projection-based cross-mapping methods and multilingual embeddings may not alter significantly the global structure of the embeddings, primarily based on semantic similarity. As a result, the transfer of cross-lingual information might be relatively low even when the degeneration issue is bypassed. Still, projection-based cross-lingual methods showcase a slight but constant improvement for distant language pairs, which coincidentally should be the cases where the alignment has a larger impact on the global structure on the embeddings. As theorized in section 4.3, it might also be possible that the implicit projection learned by the neural network is limited for language pairs of low similarity if no strong initialization is given. This is in line with previous claims on the importance of initialization in attention-based encoder-decoder translation models (Devlin et al., 2019; M. Lewis et al., 2020; Liu et al., 2020), where neural language models are especially effective.



(a) Independent monolingual embeddings.

(b) Embeddings cross-mapped using MUSE.

Figure 6.2: Exemplification of the behavior of projection-based cross-mapping using a set of translation pairs for the de-en language pairs, shown as orange and blue dots respectively. The top figure depicts the base embeddings, where no explicit alignment exists, while the bottom one corresponds with a projection of the German embedding into the English monolingual space using the MUSE method. It can be seen that for the anchor word "aspirin" the alignment is very strong, as well as for a standard pair such as "government"-"regierung", while others words the projection is not very effective. "Aspirin", "baby" and "zeitgeist" are anchor words that are written identically in both languages. On the other hand, "polizei"-"police", "newspaper"-"zeitung", "government"-"regierung" and "science"-"wissenschaft" are translation pairs that are written differently. The embeddings have been reduced to 2 dimensions using principal component analysis (PCA), with 2 components, perplexity equal to 30 and 3,500 iterations.

Chapter 7
Conclusions

Weakly-supervised cross-lingual mapping methods have develop rapidly in the recent past, showcasing an impressive performance even in low-resource scenarios and for distant language pair alignment. However, the typological and phylogenetic features that play a role in cross-mapping are yet to be fully understood. Moreover, transferring of cross-lingual information to NMT pre-training has not yielded satisfactory results so far. This work makes use of fully unsupervised cross-lingual models, which are generally outperformed by semi-supervised approaches but provide a particularly clear view of the inner workings of cross-lingual methods, to explore some of the factors that affect their effectiveness.

First, unsupervised cross-lingual models are studied on the basis of their performance in bilingual lexicon induction. The experimental results showcase agreement with prior publications regarding the usefulness of subword information in cross-linguality, and suggest that character-level encoding might be especially relevant for language pairs of low similarity. Moreover, they not only reveal that phylogenetic relatedness should not be directly taken to dictate cross-lingual performance, but also that purely semantic similarity is not the only typological feature captured by projection-based mappings. Additionally, there exist typological features that multilingual embeddings take advantage of, but that explicit cross-lingual methods have been unable to use so far.

Cross-lingual pre-training is then fully realized by integrating the cross-lingual models as pre-training in an attention-based encoder-decoder architecture. Cross-lingual transfer remains ineffective even if the structure of the pre-trained encoder and decoder is fixed during fine-tuning, which is indicative of the degeneration of cross-lingual projections in supervised training, and the limited scope of pre-trained embeddings. Modern cross-lingual approaches use joint optimization considering both an objective that ensures the structure integrity of the shared representation space and another one that maximizes the fine-tuning task.

This work hopes to provide some resources that can help our currently limited understanding of the impact that linguistic characteristics, as well model features such as subword encoding and vector space structure, have on cross-linguality and language representation as a whole. Future research may take on this body of information to design strategies to adapt cross-mapping methods to different language pairs according to their features, as well as to improve said cross-lingual techniques so that they are able to capture typological information that is currently lost. By integrating it with current pre-training approaches that rely on joint optimization, weakly-supervised cross-linguality could become the definitive pre-training strategy, and bridge some of the notable gaps that current machine translation models present when dealing with low-resource language pairs.

Appendix

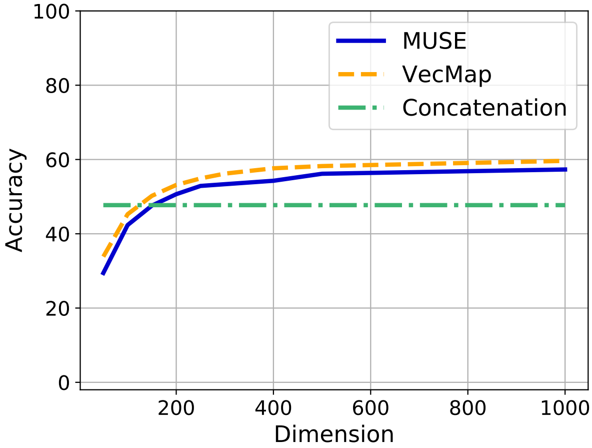


Figure 7.1: Evolution of BLI performance for each language pair and cross-mapping method according to the dimension of the monolingual embeddings for the language pair fr-en.

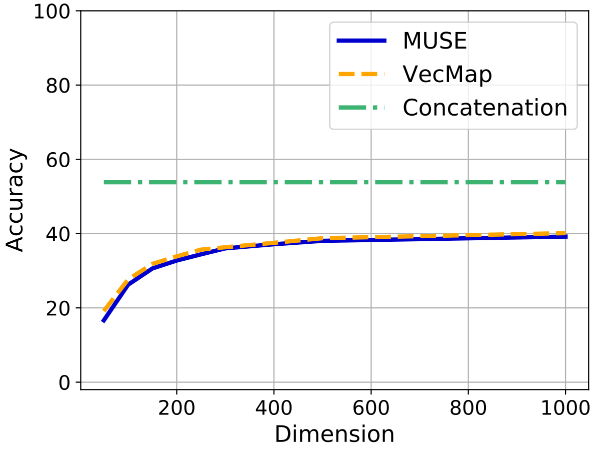


Figure 7.2: Evolution of BLI performance for each language pair and cross-mapping method according to the dimension of the monolingual embeddings for the language pair de-en.

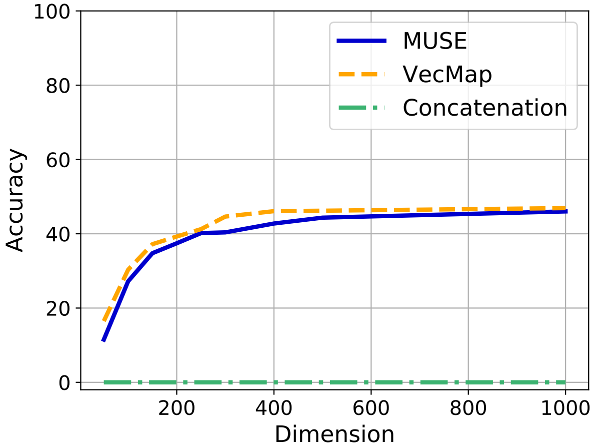


Figure 7.3: Evolution of BLI performance for each language pair and cross-mapping method according to the dimension of the monolingual embeddings for the language pair ru-en.

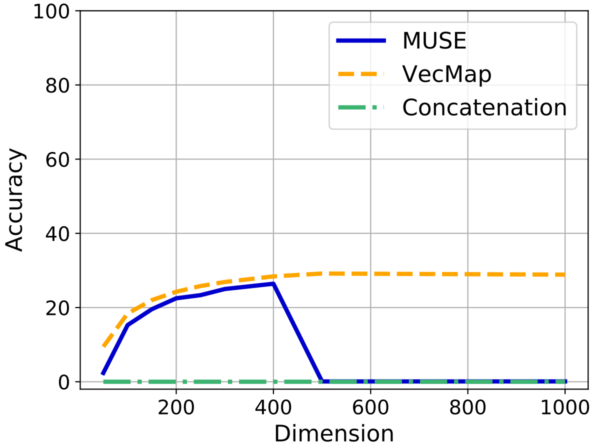


Figure 7.4: Evolution of BLI performance for each language pair and cross-mapping method according to the dimension of the monolingual embeddings for the language pair hi-en.

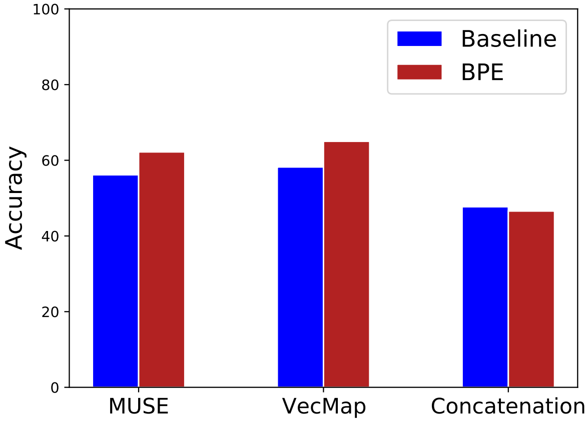


Figure 7.5: BLI performance comparison for between basic tokenization ("Baseline" in the figure) and additionally applying BPE tokenization ("BPE" in the figure) for the language pair fr-en.

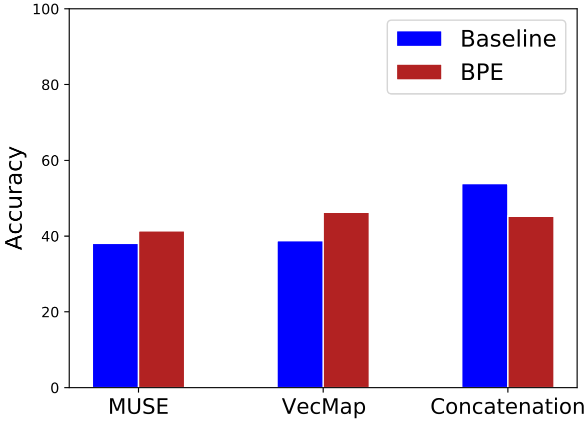


Figure 7.6: BLI performance comparison for between basic tokenization ("Baseline" in the figure) and additionally applying BPE tokenization ("BPE" in the figure) for the language pair de-en.

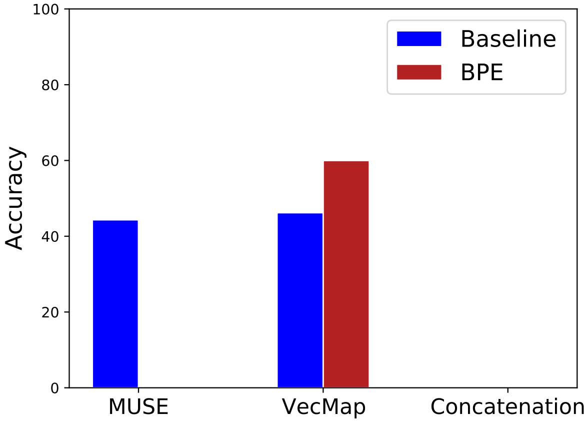


Figure 7.7: BLI performance comparison for between basic tokenization ("Baseline" in the figure) and additionally applying BPE tokenization ("BPE" in the figure) for the language pair ru-en.

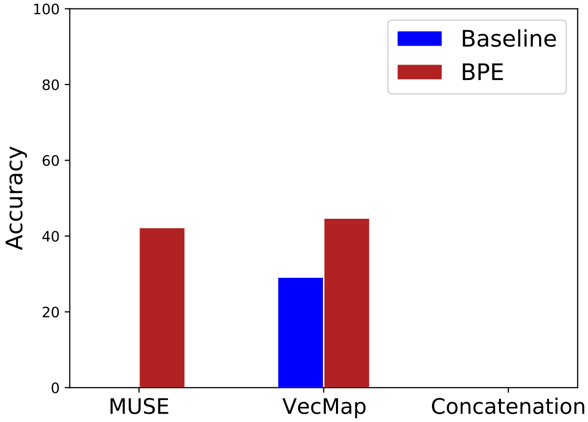


Figure 7.8: BLI performance comparison for between basic tokenization ("Baseline" in the figure) and additionally applying BPE tokenization ("BPE" in the figure) for the language pair hi-en.

Bibliography

[1] Waleed Ammar, George Mulcaire, Miguel Ballesteros, et al. “Many Languages, One Parser”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 431–444. DOI: 10.1162/tacl_a_00109. URL: https://aclanthology.org/Q16-1031.

[2] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, et al. “Massively Multilingual Word Embeddings”. In: *CoRR* abs/1602.01925 (2016). arXiv: 1602.01925. URL: http://arxiv.org/abs/1602.01925.

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 789–798.

[4] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “Bilingual Lexicon Induction through Unsupervised Machine Translation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5002–5007. DOI: 10.18653/v1/P19-1494. URL: https://www.aclweb.org/anthology/P19-1494.

[5] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “Learning principled bilingual mappings of word embeddings while preserving monolingual invariance”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2289–2294. DOI: 10.18653/v1/D16-1250. URL: https://aclanthology.org/D16-1250.

[6] Mikel Artetxe, Sebastian Ruder, et al. “A Call for More Rigor in Unsupervised Cross-lingual Learning”. In: *CoRR* abs/2004.14958 (2020). arXiv: 2004.14958. URL: https://arxiv.org/abs/2004.14958.

[7] Mikel Artetxe and Holger Schwenk. “Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3197–3203. DOI: 10.18653/v1/P19-1309. URL: https://aclanthology.org/P19-1309.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2014. arXiv: 1409.0473 [cs.CL].

[9] Amir Bakarov, Roman Suvorov, and Ilya Sochenkov. “The Limitations of Cross-language Word Embeddings Evaluation”. In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 94–100. DOI: 10.18653/v1/S18-2010. URL: https://aclanthology.org/S18-2010.

[10] Antonio Valerio Miceli Barone. “Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders”. In: *CoRR* abs/1608.02996 (2016). arXiv: 1608.02996. URL: http://arxiv.org/abs/1608.02996.

[11] Anthony J. Bell and Terrence J. Sejnowski. “The ‘independent components’ of natural scenes are edge filters”. In: *Vision Research* 37.23 (1997), pp. 3327–38.

[12] Yoshua Bengio et al. “A Neural Probabilistic Language Model”. In: 3.null (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435.

[13] Piotr Bojanowski et al. “Enriching Word Vectors with Subword Information”. In: (July 2016). URL: http://arxiv.org/abs/1607.04606.

[14] Ondřej Bojar et al. *HindMonoCorp 0.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 2014. URL: http://hdl.handle.net/11858/00-097C-0000-0023-6260-A.

[15] Kaj Bostrom and Greg Durrett. “Byte Pair Encoding is Suboptimal for Language Model Pretraining”. In: *CoRR* abs/2004.03720 (2020). arXiv: 2004.03720. URL: https://arxiv.org/abs/2004.03720.

[16] Garland Cannon. “Historical change and English word-formation : recent vocabulary”. In: *Language* 65 (1989), p. 880.

[17] Ronan Collobert and Jason Weston. “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In: Jan. 2008, pp. 160–167. DOI: 10.1145/1390156.1390177.

[18] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, et al. “Word Translation Without Parallel Data”. In: *arXiv preprint arXiv:1710.04087* (2017).

[19] Alexis Conneau, Guillaume Lample, Ruty Rinott, et al. “XNLI: Evaluating Cross-lingual Sentence Representations”. In: *CoRR* abs/1809.05053 (2018). arXiv: 1809.05053. URL: http://arxiv.org/abs/1809.05053.

[20] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[21] Mattia Di Gangi, Matteo Negri, and Marco Turchi. “Adapting Transformer to End-to-End Spoken Language Translation”. In: *INTERSPEECH*. Sept. 2019, pp. 1133–1137. DOI: 10.21437/Interspeech.2019-3045.

[22] Yerai Doval et al. “On the Robustness of Unsupervised and Semi-supervised Cross-lingual Word Embedding Learning”. In: *CoRR* abs/1908.07742 (2019). arXiv: 1908.07742. URL: http://arxiv.org/abs/1908.07742.

[23] Long Duong et al. “Learning Crosslingual Word Embeddings without Bilingual Corpora”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1285–1295. DOI: 10.18653/v1/D16-1136. URL: https://aclanthology.org/D16-1136.

[24] Manaal Faruqui and Chris Dyer. “Improving Vector Space Word Representations Using Multilingual Correlation”. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 462–471. DOI: 10.3115/v1/E14-1049. URL: https://aclanthology.org/E14-1049.

[25] Ryan Georgi, Fei Xia, and William Lewis. “Comparing Language Similarity across Genetic and Typologically-Based Groupings”. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, Aug. 2010, pp. 385–393. URL: https://aclanthology.org/C10-1044.

[26] Goran Glavaš et al. “How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 710–721. DOI: 10.18653/v1/P19-1070. URL: https://www.aclweb.org/anthology/P19-1070.

[27] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[28] Stephan Gouws, Yoshua Bengio, and Greg Corrado. “BilBOWA: Fast Bilingual Distributed Representations without Word Alignments”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 748–756. URL: https://proceedings.mlr.press/v37/gouws15.html.

[29] Stephan Gouws and Anders Søgaard. “Simple task-specific bilingual word embeddings”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 1386–1390. DOI: 10.3115/v1/N15-1157. URL: https://aclanthology.org/N15-1157.

[30] Karl Moritz Hermann and Phil Blunsom. “Multilingual Distributed Representations without Word Alignment”. In: *Proceedings of ICLR*. Apr. 2014. URL: `http://arxiv.org/abs/1312.6173`.

[31] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. “Distributed Representations”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986, pp. 77–109. ISBN: 026268053X.

[32] W. John Hutchins, Leon Dostert, and Paul Garvin. “The Georgetown-I.B.M. experiment”. In: *In*. John Wiley & Sons, 1955, pp. 124–135.

[33] Hirofumi Inaguma et al. “Multilingual End-to-End Speech Translation”. In: *CoRR* abs/1910.00254 (2019). arXiv: `1910.00254`. URL: `http://arxiv.org/abs/1910.00254`.

[34] Inigo Jauregi Unanue et al. “English-Basque Statistical and Neural Machine Translation”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. URL: `https://aclanthology.org/L18-1141`.

[35] Baijun Ji et al. “Cross-Lingual Pre-Training Based Transfer for Zero-Shot Neural Machine Translation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.01 (Apr. 2020), pp. 115–122. DOI: `10.1609/aaai.v34i01.5341`. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/5341`.

[36] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2015).

[37] Guillaume Klein et al. “OpenNMT: Open-Source Toolkit for Neural Machine Translation”. In: *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 67–72. URL: `https://www.aclweb.org/anthology/P17-4012`.

[38] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. “Inducing Crosslingual Distributed Representations of Words”. In: *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 1459–1474. URL: `https://aclanthology.org/C12-1089`.

[39] Tomás Kociský, Karl Moritz Hermann, and Phil Blunsom. “Learning Bilingual Word Representations by Marginalizing Alignments”. In: *CoRR* abs/1405.0947 (2014). arXiv: `1405.0947`. URL: `http://arxiv.org/abs/1405.0947`.

[40] Philipp Koehn, Hieu Hoang, et al. “Moses: Open Source Toolkit for Statistical Machine Translation”. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 177–180. URL: `https://aclanthology.org/P07-2045`.

[41] Philipp Koehn and Rebecca Knowles. "Six Challenges for Neural Machine Translation". In: *CoRR* abs/1706.03872 (2017). arXiv: 1706.03872. URL: http://arxiv.org/abs/1706.03872.

[42] Taku Kudo. "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates". In: *CoRR* abs/1804.10959 (2018). arXiv: 1804.10959. URL: http://arxiv.org/abs/1804.10959.

[43] Guillaume Lample and Alexis Conneau. "Cross-lingual Language Model Pretraining". In: *CoRR* abs/1901.07291 (2019). arXiv: 1901.07291. URL: http://arxiv.org/abs/1901.07291.

[44] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. "Unsupervised Machine Translation Using Monolingual Corpora Only". In: *CoRR* abs/1711.00043 (2017). arXiv: 1711.00043. URL: http://arxiv.org/abs/1711.00043.

[45] Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. "Learning Multilingual Word Representations using a Bag-of-Words Autoencoder". In: *CoRR* abs/1401.1803 (2014). arXiv: 1401.1803. URL: http://arxiv.org/abs/1401.1803.

[46] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. "Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 270–280. DOI: 10.3115/v1/P15-1027. URL: https://www.aclweb.org/anthology/P15-1027.

[47] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703.

[48] Zehui Lin et al. *Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information*. 2021. arXiv: 2010.03142 [cs.CL].

[49] Yinhan Liu et al. "Multilingual Denoising Pre-training for Neural Machine Translation". In: *CoRR* abs/2001.08210 (2020). arXiv: 2001.08210. URL: https://arxiv.org/abs/2001.08210.

[50] Thang Luong, Hieu Pham, and Christopher D. Manning. "Bilingual Word Representations with Monolingual Quality in Mind". In: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Denver, Colorado: Association for Computational Linguistics, June 2015, pp. 151–159. DOI: 10.3115/v1/W15-1521. URL: https://aclanthology.org/W15-1521.

[51] Matouš Macháček and Ondřej Bojar. “Results of the WMT14 Metrics Shared Task”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation.* Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 293–301. DOI: 10.3115/v1/W14-3336. URL: https://aclanthology.org/W14-3336.

[52] Tomas Mikolov, Kai Chen, et al. *Efficient Estimation of Word Representations in Vector Space.* 2013. arXiv: 1301.3781 [cs.CL].

[53] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. “Exploiting Similarities among Languages for Machine Translation”. In: (Sept. 2013). URL: http://arxiv.org/abs/1309.4168.

[54] Tomas Mikolov, Ilya Sutskever, et al. “Distributed Representations of Words and Phrases and their Compositionality”. In: *CoRR* abs/1310.4546 (2013). arXiv: 1310.4546. URL: http://arxiv.org/abs/1310.4546.

[55] Ndapa Nakashole and Raphael Flauger. “Characterizing Departures from Linearity in Word Translation”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 221–227. DOI: 10.18653/v1/P18-2036. URL: https://aclanthology.org/P18-2036.

[56] Usman Naseem et al. “A Comprehensive Survey on Word Representation Models: From Classical to State-Of-The-Art Word Representation Language Models”. In: *CoRR* abs/2010.15036 (2020). arXiv: 2010.15036. URL: https://arxiv.org/abs/2010.15036.

[57] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *CoRR* abs/1912.01703 (2019). arXiv: 1912.01703. URL: http://arxiv.org/abs/1912.01703.

[58] Barun Patra et al. “Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Florence, Italy: Association for Computational Linguistics, July 2019. DOI: 10.18653/v1/P19-1018. URL: https://aclanthology.org/P19-1018.

[59] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

[60] Matt Post. “A Call for Clarity in Reporting BLEU Scores”. In: *Proceedings of the Third Conference on Machine Translation: Research Papers.* Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. DOI: 10.18653/v1/W18-6319. URL: https://aclanthology.org/W18-6319.

[61] Daniel Povey et al. “The Kaldi speech recognition toolkit”. In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (Jan. 2011).

[62] T. Pyles and J. Algeo. *The Origins and Development of the English Language*. Harcourt Brace Jovanovich College Publishers, 1993. ISBN: 9780155001688. URL: `https://books.google.es/books?id=WVEdAQAAIAAJ`.

[63] Ye Qi et al. "When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation?" In: *CoRR* abs/1804.06323 (2018). arXiv: `1804.06323`. URL: `http://arxiv.org/abs/1804.06323`.

[64] Alec Radford and Karthik Narasimhan. "Improving Language Understanding by Generative Pre-Training". In: 2018.

[65] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data". In: *J. Mach. Learn. Res.* 11 (Dec. 2010), pp. 2487–2531. ISSN: 1532-4435.

[66] Shuo Ren et al. "Explicit Cross-lingual Pre-training for Unsupervised Machine Translation". In: *CoRR* abs/1909.00180 (2019). arXiv: `1909.00180`. URL: `http://arxiv.org/abs/1909.00180`.

[67] Sebastian Ruder. "A survey of cross-lingual embedding models". In: *CoRR* abs/1706.04902 (2017). arXiv: `1706.04902`. URL: `http://arxiv.org/abs/1706.04902`.

[68] P. Schönemann. "A generalized solution of the orthogonal procrustes problem". In: *Psychometrika* 31 (1966), pp. 1–10.

[69] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *CoRR* abs/1508.07909 (2015). arXiv: `1508.07909`. URL: `http://arxiv.org/abs/1508.07909`.

[70] Or Sharir, Barak Peleg, and Yoav Shoham. "The Cost of Training NLP Models: A Concise Overview". In: *CoRR* abs/2004.08900 (2020). arXiv: `2004.08900`. URL: `https://arxiv.org/abs/2004.08900`.

[71] Samuel L. Smith et al. "Offline bilingual word vectors, orthogonal transformations and the inverted softmax". In: *CoRR* abs/1702.03859 (2017). arXiv: `1702.03859`. URL: `http://arxiv.org/abs/1702.03859`.

[72] Anders Søgaard, Sebastian Ruder, and Ivan Vulić. "On the Limitations of Unsupervised Bilingual Dictionary Induction". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 778–788. DOI: `10.18653/v1/P18-1072`. URL: `https://www.aclweb.org/anthology/P18-1072`.

[73] Matthias Sperber et al. "Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation". In: *CoRR* abs/1904.07209 (2019). arXiv: `1904.07209`. URL: `http://arxiv.org/abs/1904.07209`.

[74] Fred Stentiford and M.G. Steer. "Machine Translation of Speech". In: *British Telecom Technology Journal* 6 (Apr. 1988), pp. 116–123.

[75] Haipeng Sun et al. "Unsupervised Bilingual Word Embedding Agreement for Unsupervised Neural Machine Translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1235–1245. DOI: `10.18653/v1/P19-1119`. URL: `https://aclanthology.org/P19-1119`.

[76] Wilson L. Taylor. ""Cloze Procedure": A New Tool for Measuring Readability". In: *Journalism Quarterly* 30.4 (1953), pp. 415–433. DOI: `10.1177/107769905303000401`. eprint: `https://doi.org/10.1177/107769905303000401`. URL: `https://doi.org/10.1177/107769905303000401`.

[77] A.M. Turing. "I.—Computing Machinery and Intelligence". In: *Mind* LIX.236 (Oct. 1950), pp. 433–460. ISSN: 0026-4423. DOI: `10.1093/mind/LIX.236.433`. eprint: `https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf`. URL: `https://doi.org/10.1093/mind/LIX.236.433`.

[78] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: `1706.03762`. URL: `http://arxiv.org/abs/1706.03762`.

[79] Ivan Vulic, Goran Glavas, et al. "Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?" In: *CoRR* abs/1909.01638 (2019). arXiv: `1909.01638`. URL: `http://arxiv.org/abs/1909.01638`.

[80] Ivan Vulic and Marie-Francine Moens. "Bilingual Distributed Word Representations from Document-Aligned Comparable Data". In: *CoRR* abs/1509.07308 (2015). arXiv: `1509.07308`. URL: `http://arxiv.org/abs/1509.07308`.

[81] Ivan Vulić and Anna Korhonen. "On the Role of Seed Lexicons in Learning Bilingual Word Embeddings". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 247–257. DOI: `10.18653/v1/P16-1024`. URL: `https://aclanthology.org/P16-1024`.

[82] A. Waibel et al. "JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies". In: *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing* (1991), 793–796 vol.2.

[83] Rui Wang and Hai Zhao. "Advances and Challenges in Unsupervised Neural Machine Translation". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. online: Association for Computational Linguistics, Apr. 2021, pp. 17–21. URL: `https://aclanthology.org/2021.eacl-tutorials.5`.

[84] Min Xiao and Yuhong Guo. "Distributed Word Representation Learning for Cross-Lingual Dependency Parsing". In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2014, pp. 119–129. DOI: `10.3115/v1/W14-1613`. URL: `https://aclanthology.org/W14-1613`.

[85] Chao Xing et al. "Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 1006–1011. DOI: `10.3115/v1/N15-1104`. URL: `https://aclanthology.org/N15-1104`.

[86] Jiacheng Yang et al. "Towards Making the Most of BERT in Neural Machine Translation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 9378–9385. DOI: `10.1609/aaai.v34i05.6479`. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/6479`.

[87] George K. Zipf. "Human behavior and the principle of least effort. Cambridge, (Mass.): Addison-Wesley, 1949, pp. 573". In: *Journal of Clinical Psychology* 6.3 (1950), pp. 306–306. DOI: `https://doi.org/10.1002/1097-4679(195007)6:3<306::AID-JCLP2270060331>3.0.CO;2-7`.

[88] Will Y. Zou et al. "Bilingual Word Embeddings for Phrase-Based Machine Translation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1393–1398. URL: `https://aclanthology.org/D13-1141`.

List of Figures

List of Tables